Analyzing Difficulty Index of Ekadanta English Section Multiple Choice

Dara Yusnida[1✉], Nour Ayouni[2]
[1]*English Language Department, Universitas Iskandarmuda, Aceh, Indonesia,*
[2]*English Language Department, Universitas Iskandarmuda, Aceh, Indonesia*
✉ email: dara_yusnida@yahoo.co.id

ABSTRACT

Evaluation has been an essential component of the teaching and learning process for as long as it has been a part of the teaching and learning process itself, which is something that takes place all the time at every educational institution. This is because the assessment has a tight link with educators in the process of transferring information to analyze how much students' abilities have developed over the course of the learning process and while assembling each assignment in the process of learning. This study tries to shed light on the index of difficulty index found in a set of prediction test in Ekadanta tutoring book. This study employed descriptive qualitative approach. The data source was 20 questions from Ekadanta tutoring book. The subject of this study were 11 students of English Department at Universitas Iskandarmuda, Banda Aceh, Indonesia. The data collection was carried out by instructing the students to do the test within 40 minutes. After the data obtained, they were analyzed using index of difficulty formula and categorized into its index level. The results shows that from 20 question, there are 8 questions fall under difficult index, 8 questions are categorized into moderate, and the other 4 questions are considered easy in the level of difficulty index. It is concluded that the test is good and descent from the perspective of the difficulty index. It is later suggested that future research would involve more questions and test sets as well as participants in studies alike.

Keywords: *Evaluation; Index Difficulty; Language Testing; Language Test; Test Item Analysis.*

INTRODUCTION

Evaluation is the topic that receives the greatest attention from knowledgeable educators and academic advisers, and it is discussed in relation to all aspects of the teaching and learning process within the context of a school setting. The assessment of the learning process is accorded a particular focus in each and every research endeavor that is carried out. This is due to the fact that evaluation has a significant impact on the process of quality improvement, which, when applied to the teaching system, is expected to result in an instructional method that is both successful and efficient. It is common knowledge that assessment in the field of education has a very wide breadth, with at the absolute least the need that it encompasses three fundamental objectives that one should constantly strive to accomplish. These goals, which are known as the process, the program, and the teaching, are the outcomes of the evaluation program itself. When

evaluating the educational process, it typically characterizes the process, as well as the organization of the stages of the teaching and learning process, in relation to the activities that involve contact between educators and participants in the learning process. The evaluation of the entire teaching component in an educational institution is referred to as the program evaluation. This evaluation encompasses not only the goals of the education, but also the content, the structure of the teaching and learning process, and even the goals that are intended to be achieved from the education program itself. The evaluation of learning outcomes, also known as educational outcomes, looks at how time is allocated, and plans are made for programs that will be put into action both in the near term and in the longer term. The evaluation of educational outcomes includes, among other things, an assessment procedure in the cognitive, emotional, and psychomotor domains of a student's development.

The importance of evaluation has set the alignment to be achieved in his study. Since this circumstance has a highly positive impact on educators, helping them to organize and strategize effective and efficient learning procedures and strategies for the future (Prasetyo, 2018). Evaluation is also a significant consideration in the final decision-making process because it is part of a systematic method that gathers information for all of the teaching staff. Evaluation has been an essential component of the teaching and learning process for as long as it has been a part of the teaching and learning process itself, which is something that takes place all the time at every educational institution. This circumstance has a very positive impact on educators in terms of assisting them in organizing and strategizing effective and efficient learning procedures and strategies for the future (Prasetyo, 2018). Evaluation is also a significant consideration in the final decision-making process because it is part of a systematic method that gathers information for all of the teaching staff.

In general, in order for educators to improve the quality of a question or exam, they are necessary to review each of the questions that are already in use. This procedure can serve as a standard for summarizing and utilizing information in order to enable educators to make informed judgments on each and every assessment that is currently in use (Deyger & Gorp, 2015). It is not difficult to find educators who are capable of formulating strategies for effective and efficient teaching and learning processes. However, the ability of educators to formulate strategies for effective and efficient teaching and learning processes is not the same as their ability to apply these strategies in the context of the teaching and learning process. This explains why it is essential as well as vital for all professional educators to have a high degree of competence and talent in order to accomplish the goals that have been set forth for the program. The process of improving the abilities and skills of teaching staff is not a process that can be mastered in a short period of time. However, over the course of time, there is a great possibility for all teaching staff members to have good abilities and skills, which will allow them to become competent professional teachers. Every single evaluation strategy always contains learning results that may be utilized to establish valid and reliable standards for following research procedures and continue for other new research investigations. These learning outcomes can also be used to inform future evaluation strategies. It is necessary

to complete two primary stages in order to generate valid and trustworthy benchmarks. These two stages are the process of designing good measuring instruments and the method of analyzing these measuring instruments.

Specifically speaking under EFL circumstances, English is regarded by educators as a learning topic that contributes to and has full responsibility for the overall improvement of the quality of pupils because of its status as an international language. This is due to the fact that English is a required subject in the Indonesian national curriculum, despite the fact that it is categorized as a foreign language that originates in another nation. In spite of the fact that English is a foreign language for people who come from other countries, the process of teaching English still has a primary aim, and it is assumed that every student will be able to accomplish this purpose. To ensure that the purpose is well-achieved, teachers and educators are obliged to carry out an evaluation in the process of measuring students' skills in the subject area of English in order to ensure that the goals that have been established in the current program of evaluation are met.

In the context of a teaching and learning process, the quality of a test that is deemed to be good has the potential to affect and develop the abilities of both students and the teaching staff members themselves. This is necessary to achieve findings that are valid and trustworthy (Harding & McNamara, 2017). In the process of designing measurement tests that will later be given to students, educators are not permitted to arrange questions according to their own wishes. This is done to ensure that the designed learning program that has been prepared is not violated or is not in conflict with any of the questions on the measurement test. Tests can be an alternate tool that is ideal for use in order to attain these aims. They can help improve students' performance as well as the quality of their work and their excitement for the learning process (Heaton, 1989).

Validity, practicality, reliability, and analysis are the four aspects of a measurement test that must be present for the test to be considered of high quality (Heaton, 1989). A measurement test must possess all four of these aspects in order to be considered of high quality. The term "validity" refers to a link that can be drawn between the contents of each item on a test and the objective of the examination that is being conducted. Every question on the exam needs to adhere to certain practical requirements in order to ensure that students are able to comprehend the material and do well on the test. The purpose of the reliability criteria is to explain how the test should behave in order for it to provide the same result regardless of whether it is used in research on one group of students at one time or on many groups of students at different times (Grapin, 2022). In conclusion, because of the criteria for the study, each individual item in the test comprised a unique combination of strengths and challenges.

In general, all teachers do evaluate the abilities of their students by administering tests. The purpose of these evaluations is to determine how far their students' abilities have progressed. The type of test that is utilized is one that is evaluated based on the number of scores obtained, such as by employing a test with a model consisting of multiple-choice questions (Grapin, 2022). In the context of employing the model of multiple-choice questions as an assessment tool connected to how much students'

capabilities have progressed. When designing these questions, there are two aspects that need to be taken into consideration. First, the questions need to be able to measure the construct against the target that is being measured, and second, the questions need to be uni-dimensional (meaning that they do not differentiate between one student, and another based on the level of the student) (North & Piccardo, 2023). The preceding description makes it abundantly clear that when using the multiple-choice question model; it is necessary to perform an analysis in advance concerning the questions that will be given to students to determine whether or not these questions can provide valid measurement results in order to eliminate scoring anomalies. This is done in order to avoid scoring discrepancies.

Here emerges the potential problem formulated in this current study. Mostly, teachers, tutors, or educators rarely conduct any evaluation of the tests they have designed. The test has to be analyzed at each item level to consider its index of difficulty, discriminant power, and distractor effectiveness. The process of assessing each question item is referred to as item analysis. It is important for educators to evaluate correctly if each question item has the power to make educators aware of the strengths and shortcomings of pupils (Toughiry & Ziafar, 2014). In addition, there is another process that takes place, and that process is the elaboration process of explaining the capacity of the exam to classify students depending on how effectively students perform the assignment (Saragih, 2017). The sole objective of the research analysis that is performed is to determine which questions should be retained for use in subsequent evaluations and which should be discarded. This is done with the intention of achieving positive findings (Susanti et al., 2022).

Language Testing

Various experts have various perspectives on the essence of language testing. According to Heaton (1989), a Proficiency Test may be used to assess students' abilities in assessing how well they grasp a foreign language that is studied in relation to future usage of that language. Where the aim of the test is not to assess students' understanding of the language, but rather how well they perform on the test using the established criteria. Wang & Li (2020) refers to the following tests: (1) Summative exams, which assess how far a student has progressed in a learning process. Educators have a lot of procedures in developing student accomplishment by employing two types of assessments: progress achievement tests (provided during the learning process every semester) and final achievement tests (issued at the conclusion of the learning process). (2) Diagnostic tests are tests that provide clear analysis findings of students' strengths and shortcomings in each topic; this exam is often provided by teachers when educators want to know the next effective and efficient teaching material to increase learners' skills. (3) The Placement Test is intended to assist the teacher in classifying students' abilities based on overall ability; in general, this test is given by a teacher at the beginning of the period when the learning process begins so that each student is divided into groups with equal academic abilities for the purpose of preventing gaps and inequalities of abilities so that the learning process can run effectively and efficiently; (4) Direct and indirect

tests. The indirect exam, on the other hand, is a test that asks students to perform what the teacher instructs, which the teacher believes to be a fundamental guideline for each student who is characterized as having difficulties overcoming some of the abilities connected to the information being assessed. According to Wang & Li (2020), this exam asks educators to assess abilities such as the ability to write without having to write. (5) Norms-reference testing demands educators to see and carefully analyze the level of ability of each student as a whole before constructing a test that will be administered and carried out by students (Jerrold, 2012). The criterion-reference test, on the other hand, is intended to assess students in the process of interacting socially with both fellow students and teachers, as well as in expressing and soliciting opinions based on the correct discussion structure, rather than on consideration of what each participant is capable of doing. And the last, (6) Objective and subjective tests, which are only separated by the method of assessment, are ideal for evaluating learning outcomes because they employ a higher number of questions that may cover wide learning contents with easy assessment methodologies and have a high level of dependability (dos-Santos & Ramírez-ávila, 2023). The subjective test, on the other hand, is a test that is less efficient in measuring knowledge but very good in measuring the level of synthesis and evaluation because it uses relatively few questions and only covers a portion of the material asked, and the method of assessment is also classified as difficult and less reliable due to the presence of scores. Each question is subjective. As a result, objective testing will be better if the reliability value is higher, and subjective test findings will be better or more legitimate if the subjectivity value on this test is lower (Harding & McNamara, 2017).

Reading and reading and writing in tests are seen differently from their other counterparts. Brown (2004) describes that the formation of a good reading test, it is required educators provide pieces of a foreign language with different expressions and lengths, add distinctions between graphemes and orthography, determine important words from the reading material, sort words according to the correct language structure, providing good grammatical, providing meaning that can be expressed in different parables but still having the same meaning, and presenting a cohesive device in a written discourse. There are numerous techniques for reading and writing exams that may be employed, including textual reading and writing, cognitive reading and writing, and contextual reading and writing (Khatimah et al., 2022). According to Hyland (2022), every teacher is expected to forsake the traditional technique in producing a reading and writing exam, and it is also advised that every teacher views reading and writing as a tool for communication rather than a learning duty to know how to organize effective reading and writing. Reading and writing is a socio-cultural action that is difficult to analyze as a whole, even when employing evaluations based on theoretical sources because reading and writing are not simply reliant on the situational environment of the writer's individual cognitive processes. As a result, the evaluation necessitates the use of a rubric or a socio-cognitive assessment paradigm.

As previously said, the reading and writing process is greatly impacted by the writer's area of expertise and memory abilities, as reading and writing are socio-cultural in nature. As a result, as the primary criterion, the writer must have meticulously

prepared objectives, which include outlining rhetorical concerns that will be studied, assessed, amended, and eventually summed into a written work. If inadequacies in the findings of the analysis are discovered, the process will be re-evaluated in the next review, and each process will always be controlled by an executive control, also known as a monitor.

More attention and understanding are required in reading and writing under certain conditions so that the writer can provide discussion results that are in accordance with what is expected and in accordance with the reader's interests by using language appropriate to the reader's discourse (Hyland, 2022). As with reading and writing English as a conditioned topic, the writer is expected to be accountable; establish a clear link between the author and the outcomes of the author's description, and to arrange his works in the right sequence. In summation, both the writer and the examiner must have suitable involvement with the issue under consideration in order to make effective use of the degree of language expertise, particularly linguistics, discourse, and sociolinguistics (Ismail & Yoestara, 2022).

Difficulty Index

In general, when determining the worth of learning in a research population that is both relatively large in number and diverse in nature, two population groupings will be found (Heaton, 1989). These population groupings are referred to as the upper group (the control group; the group with high scores) and the lower group (the treatment group; the group with low scores). The purpose of this grouping is to determine the population's average value, and then review how much the lower group increased in comparison to the upper group. The norm-based method, often known as the group assessment method, is another name for this technique. In the process of assessing the research, if it is discovered that the results of the research form an asymmetrical curve, either by forming a slanted line to the right or vice versa, the researcher is required to re-analyze each question item that was used as a measuring instrument in the study. This is because the utilization of this measuring instrument determines the value of the research's success. The purpose of item difficulty analysis is to determine whether or not the measuring equipment is appropriate and satisfies the standards that have been given. It is hoped that the researcher will be able to find information that can later be used as reference material to later be used as a benchmark in compiling new question items or replacing question items that do not match the existing criteria from each of the results of the analysis that was carried out. After going over these results, they will be reevaluated to determine which things are suitable and legitimate based on the criteria that already exist so that they may be utilized again in this research procedure.

During the process of analysis that was discussed above, the researcher also needs to evaluate the difficulty of the existing question items. This was done with the intention of classifying each question item according to the level of complexity, which ranges from simple to complex (please see Method section for detail about the index).

Earlier studies have obtained several various results. The researchers analyzed the usage of summative tests in the field of English by employing a variety of different

approaches and tools. Amelia (2010) conducted an analysis of the degree of difficulty involved in putting into practice the Summative Test for English topics at MTs Darul Ma'arif Jakarta. The study considered the results to be "moderate" in item difficulty. As for a different study, Rosdiana & Ismail (2017) conducted a study on item analysis on questions used at an international science school in Banda Aceh. According to the findings, 84% of the items are classified as being in the "easy" index, 11% are classified as being in the "moderate" index, and 4% are classified as being in the "difficult" index. The findings also indicate the effectiveness of just 17% of the various distractors. The findings lead one to the conclusion that the formative exam items are, in fact, simple for the students of science; nevertheless, there is a catch: the majority of the distractor questions do not function appropriately for their level of cognitive ability. In the study that was carried out by Karim et al. (2021), the results were obtained by using a quantitative analysis research method and a descriptive qualitative analysis approach to each question item on the Level I Radiography training exam questions. Additionally, several different levels of difficulty were discovered on the General and Specific exam questions. In the General exam, out of a total of forty questions for each question, it was discovered that seven questions had a level of difficulty that did not match the level of difficulty that should have been programmed. Additionally, in the specific exam questions, it was discovered that twelve questions out of the total of sixty questions in the exam had a level of difficulty that did not match what was programmed. As a result of these findings, it was determined that the difficulty level of each item on the examination needed to be reevaluated before it could be used for examinations at a later date in other time periods. Furthermore, as some of the items with a differentiating level were deemed to have a very low level of difficulty, it would be preferable if they were either changed or eliminated.

This study is considered important to be carried out as it has two types of significance focusing on the theoretical field and the practical field, respectively. The results of the research that has been described above have led to the researcher having these results, and the results have been described above. In the realm of theory, the researchers have hopes that the findings of this study will become a source of knowledge that can later be applied to the process of analyzing the questions that are being formulated in order to improve the formulation of those questions. As a result, the researcher anticipates that the findings of this study will be able to serve as valid parameters in the process of seeking and determining decisions to formulate judgments and postulates based on an ontology related to the teaching of English (Fazlollahi et al., 2015). In the practical world, the researchers hope that the findings of this study will be able to assist teachers in the process of compiling and designing tests that are regarded as being good and that is in accordance with the learning goals of the curriculum. This will allow teachers to more accurately gauge the level of progress that students have made in their knowledge. The researcher hopes that this research can also be used as a comparison to determine how good the level of knowledgeability of each individual student is, and it is expected to be an additional reference in writing subsequent research relating to the study of the same topic. In addition, the researcher also hopes that this

research can be used as a comparison to determine how good the level of knowledgeability of each individual student is. Based on the rationale above, the following research question is formulated: *"What is the item difficulty index of each item in Ekadanta prediction questions in the English section?"*

METHOD

The method used in this research is the descriptive qualitative method. This method was chosen because the fact that giving precise accounts of perceptions is the most frequently cited justification for the adoption of a descriptive technique, especially in fields where little is known about the subject being researched, in this case it is about the difficulty index. The objects of this study were 20 questions taken from Ekadanta's tutoring book. Ekadanta is a tutoring program in Aceh initiated by graduates of the Bandung Technical Institute. Ekadanta is one of the school's after-hours activities that is done to give extra lessons to the students so they can get better grades or have more successful learning outcomes. Ekadanta is merely founded in Bandung by several engineering students. Then, this tutoring program has also been established in Aceh in focus to reinforce school subject in pertinent to university exam entrance for Science-Based major such as Math, Physics, Chemistry, and English. This question was then taken because it needed to be tested for the level of difficulty considering the background of the test designer(s) of the question was not within education discipline; instead, they are from engineering discipline. This question was then worked on by 11 students majoring in English education at Iskandarmuda University. The criteria for the subjects chosen are: (1) They have pre-intermediate level of English ability, and (2) They have constant contact with English in th regular basis. The questions in the Ekadanta are targeted at high school graduates—or prediction questions for university entrance exams.

During data collection, each student was given a handout containing 20 questions. Then they were given 40 minutes to work on it. The questions given are in the form of reading comprehension questions. Then, after the data is collected, the data is analyzed using the problem difficulty analysis formula and matched with the available index to determine the category of each question.

Determining the difficulty level (P value) of each question, is done by analyzing each item and adding up each correct answer per individual to later calculate it using the formula previously described. In measuring the difficulty score of each question item, the grammar and substance of each material are not included in the assessment parameters to be evaluated. After the results of the difficulty value are found, they will be re-evaluated using the categorization of difficulty and the differentiating power of each of the individual question items using the following index (Heaton, 1989):

Table 1. Index of Difficulty

| Difficult | 0.00-0.30 |
| Moderate | 0.31- 0.70 |

Easy	0.71 – 1.00

In the process of categorizing each question item with a difficulty value of 0.00 (P = 0.00) and a difficulty value of 1.00 (P = 1.00). The researcher will re-evaluate. It is intended that each question item is valid and reliable so that it can be reused in research that will be carried out in the next period (Sudijono, 2008). In the evaluation process carried out by the researcher, the researcher will continue to analyze each question item so that it remains in line with the targets to be achieved based on the program that is structured according to curriculum guidelines.

In general, this is referred to as the difficulty index number, and it is symbolized by the letter P or *percentage*. The tiny number is what defines the difficulty category of the assessment questions on the results of the learning process. If the difficulty index number is 0.00, then the question is considered to be extremely tough, and none of the students are able to provide a solution to it. The difficulty index number can vary anywhere from 0.00 to 1.00. If P equals 1.00, then the question is considered to have a difficulty level that is too easy since it is assumed that every student would be able to correctly answer the questions that are presented to them.

The following formula can be used to determine the difficulty value of a question (Arikunto, 2010):
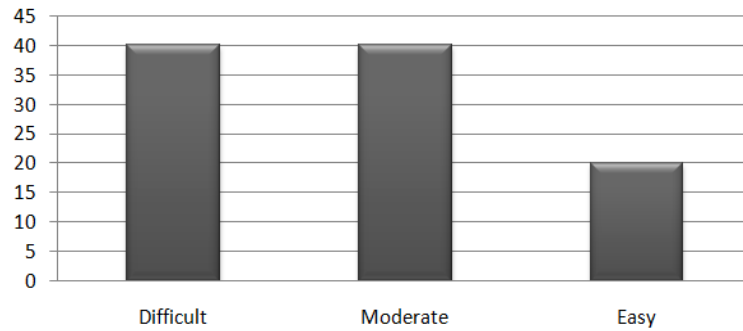
$$P = \frac{N_p}{N}$$

Where:

P = index of difficulty,

Np = Number of participants who answered the questions correctly,

N = Total number of participants who answered.

FINDINGS AND DISCUSSION

In an attempt to initially perform data analysis, the researchers separated the score into two categories, upper and lower, to derive the index of difficulty from each question item. The upper group has the highest score, while the bottom group has the lowest score. The correct answers from both groups were then computed and categorized. As mentioned earlier, the question item is considered difficult if its index is between 0.00-0.30; it is moderate—which means not too difficult nor too easy—is its index is 0.31-0.70; and lastly, the item is easy if it has the index of 0.71-1.00. The data analysis revealed that the difficulty of the question items is depicted in the graph below.

Figure 1. Recapitulative Results

From the graph, we can see that there is 40% of the items which is considered to be in **_difficult_** category; there are 8 questions. For **_moderate_** category, there is 40% of the items, which are 8 questions. And for **_easy_** category, the test has 20% of easy questions, which are 4 questions. More detailed information can be seen in table 2 below.

Table 1. Result of Difficulty Index

Index of Difficulty	Criteria	Item Number	Total
0.00 – 0.30	Difficult	11, 12, 13, 14, 16, 18, 19, 20	8
0.31 – 0.70	Moderate	4, 6, 7, 8, 9, 10, 15, 17	8
0.71 – 1.00	Easy	1, 2, 3, 5	4

The table shows that there are just four questions in the easy category. They are numbered 1, 2, 3, and 5. Question 1 has an index difficulty of 0.88. It was accurately answered by all 11 pupils. Question 2 has an index difficulty of 0.96, question 3 has an index difficulty of 0.80, and question 5 has an index difficulty of 0.99. Later, there are eight questions in the moderate category that have been found to be of moderate difficulty. Question 4, question 6, question 7, question 8, question 9, question 10, question 15, and question 17 are the ones. Question 4 has an index difficulty of 0.55. The index difficulty of Question 6 is 0.65. Next, question 7 with an index difficulty of 0.56 was successfully answered by 3 students and incorrectly answered by 8 students. Later, Question 8 has the index difficulty of 0.65. Question 9, Question 10, Question 15, and Question 17 has the index of difficulty of 044, 0.52, 0.60, and 0.54, respectively. Finally, for the difficult category, there are also question 8 questions they are, question 11, question 12, question 13, question 14, question 16, question 18, question 19, and question 20. The index difficulty of question 11 is 0.27. Question 11 was answered correctly by 1 student only and 10 students answered it wrong. The index difficulty of question 12 is 0.22. The index difficulty of question 13 is 0.27. For question 13, the index difficulty is 0.20; question 14 and question 16 has a similar index, which is 0.21. Then, for question 18 and question 19, the index difficulty is 0.29 and 0.23, respectively. Last,

the index difficulty of question 20 is 0.30. The following is presented as one of the reading texts taken from the Ekadanta tutoring book.

MAIN IDEA EXERCISE

The structure of the human brain is composed of three main parts: the forebrain, midbrain and hindbrain, each with multiple parts. and controls our complex behaviors, including senses such as vision, touch, body awareness and spatial orientation. It plays important roles in integrating sensory information from various parts of our body, knowledge of numbers and their relations, and in the manipulation of objects. Portions are involved with our visuo-spatial processing, language comprehension, the ability to construct, body positioning and movement, neglect/inattention, and left-right movements.

The Cerebrum: Also known as the cerebral cortex, the cerebrum is the largest part of the human brain, and it is associated with higher brain function such as thought and action. nerve cells make up the gray surface, which is a little thicker than our thumb. White nerve fibers beneath the surface carry signals between nerve cells in another parts of the brain and body. Its wrinkled surface increases the surface area, and is a six-lay red structure found in mammals, called the neocortex. it is divided into four sections, called differentiation and self-awareness/insights. Occipital Lobe – The occipital lobe is located at the back of our brain, and is associated with our visual processing, such as visual recognition, visual attention, spatial analysis sis (moving in a 3-D world) and visual perception of body language; such as postures, expressions and gestures. Temporal Lobe – The temporal lobe is located near our ears, and is associated with processing our perception and "lobes". they are; the frontal lobes, the parietal lobe, the occipital lobe and the temporal lobes. Recognition of auditory stimuli (including our ability to focus on one sound among many, like listening to one voice among many at a party), comprehending spoken language, verbal memory, visual memory and language production (including fluency and wordfinding),general knowledge and autobiographical memories.

A deep furrow divides the cerebrum into two halves, known as the left and right hemispheres. And, while the two hemispheres look almost symmetrical, each side seems to function differently. The right hemisphere is considered our creative side, and the left hemisphere is considered our logical side. A bundle of right hemisphere is considered our creative side, and axons, called the corpus callosum, connects the two hemispheres.

From the result of index difficulty, it can be concluded that the test items: 8 out of 20 items, are difficult. They have an index difficulty value under 0.30. Meanwhile, another 8 have a moderate index, and the other 2 items have an easy index. This means that the test has performed a good difficulty index. According to (Heaton, 1989), the index of difficulty, also known as the facility value, of an item only indicates how simple or complex the specific item that was supplied in the exam was. In most cases, it is depicted as a percentage of pupils who have the ability to provide an accurate response. It is evident from this that the students who responded with the right answer to a particular question contribute to the process of defining the difficulty level of the item. According to (Arikunto, 2010), a good test item has a specific degree of complexity that is not expected to be overly simple nor overly complex. It is anticipated that the difficulty index will fall somewhere in the moderate range. Since it is more difficult to determine a score when the item being tested is too simple, the reliability of the assessment suffers when this occurs. If the task was excessively challenging, the same thing would be perceived in exactly the same way.

CONCLUSION

In conclusion, this study confirms that, despite multiple challenges, questions designed need to be further analyzed in their conformity to language students. However, the questions are good to increase students' ability in challenging their reading and writing ability during the exercise time of the test. It is suggested that item analysis for tests is kept on being carried out because the item analysis activity is critical in the

formulation of questions in order to produce high-quality, effective item items. The advantages of item analysis include being able to determine which questions are defective or not functioning properly, improving the items through the three components of the analysis, namely, level of difficulty, discriminating power, and question distractors, and being able to revise questions that are not relevant to the material being taught, as indicated by the number of students who were unable to answer certain questions.

This study is not without limitation, a larger breadth of prediction tests and questions need to be involved in future research. In addition, more respondents are also necessary to make the findings more generalizable.

REFERENCES

Amelia, R. (2010). An analysis of english summative test in term of difficulty level (Unpublished Thesis). [UIN Syarif Hidayatullah]. https://repository.uinjkt.ac.id/dspace/handle/123456789/5888

Arikunto, S. (2010). Dasar-dasar evaluasi pendidikan. PT Bumi Aksara.

Brown, H. D. (2004). Language assessment: Principles and classroom practice. Longman.

Dos Santos, J. C., & Ramírez-ávila, M. R. (2023). Students' perspectives on the 4/3/2 technique and self-assessment to improve english speaking fluency. Studies in English Language and Education, 10(1), 41–59. https://doi.org/10.24815/siele.v10i1.25700

Fazlollahi, E., Marefat, F., & Vaezi, S. (2015). The application of dynamic assessment: Is it worth the effort? Journal of Language Teaching and Research, 6(5), 985–992. https://doi.org/10.17507/jltr.0605.10

Grapin, S. (2022). Assessment of english learners and their peers in the content areas: Expanding what "counts" as evidence of content learning. Language Assessment Quarterly, 20(2), 215–234. https://doi.org/https://doi.org/10.1080/15434303.2022.2147072

Harding, L., & Mcnamara, T. (2017). Language assessment: The challenge of elf. The Routledge Handbook of English as a Lingua Franca, February, 570–582. https://doi.org/10.4324/9781315717173-46

Heaton, J. B. (1989). Writing english language tests. Longman.

Hyland, K. (2022). Authority and invisibility: Authorial identity in academic writing. Journal of Pragmatics, 34(8), 1091–1112. https://doi.org/https://doi.org/10.1016/s0378-2166(02)00035-8

Ismail, N. M., & Yoestara, M. (2022). Validity test using principal component analysis (pca) involving research instruments on learner-centeredness, knowledge-centeredness, and assessment-centeredness. Proceedings of International Conference on Multidiciplinary Research, 5(1), 66–71. https://doi.org/10.32672/pic-mr.v5i1.5255

Jerrold, F. (2012). The roles of assessment in language teaching. English Teaching Forum, 50(3), 32.
https://files.eric.ed.gov/fulltext/ej997527.pdf

Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. Edulite: Journal of English Education, Literature and Culture, 6(2), 256–269. https://doi.org/10.30659/e.6.2.256-269

Khatimah, H., Yanelda, S., & Muslem. (2022). Delving debaters' self-development strategies in enhancing their speaking perfomance. JL3T (Journal of Linguistics, Literature and Language Teaching), 8(1), 39–47.
https://doi.org/10.32505/jl3t.v8i1.3982

North, B., & Piccardo, E. (2023). Aligning Language Frameworks: An example with the clb and cefr. Language Assessment Quarterly, 20(2), 143–165. https://doi.org/10.1080/15434303.2023.2184266

Prasetyo, A. H. (2018). Language assessment literacy development: a student-teachers experiences in teaching practice program. LLT Journal: A Journal on Language and Language Teaching, 21(2), 199–206.
https://doi.org/doi: https://doi.org/10.24071/llt.v21i2.1224

Rosdiana, R., & Ismail, N. M. (2017). Cognitive inquiry: Is english really difficult for science students? Getsempena English Education Journal, 4(1), 34–46. http://geej.stkipgetsempena.ac.id/home/article/view/50

Saragih, F. H. (2017). Testing and assessment in english language instruction. Jurnal Bahasa Unimed, 27(1), 74–79.
https://doi.org/10.24114/bhs.v27i1.5683

Sudijono, A. (2008). Pengantar evaluasi pendidikan. Raja Grafindo Persada.

Susanti, A., Widiati, U., Cahyono, B. Y., & Sharif, T. I. S. T. (2022). Assessing episodes in verbalization process of efl students' collaborative writing. Studies in English Language and Education, 9(2), 539–553.
https://doi.org/10.24815/siele.v9i2.20165

Toughiry, N., & Ziafar, M. (2014). A review article on internet-centered language assessment: Origins, challenges, and perspectives. Open Journal of Education, 2(2), 71–77.
https://doi.org/10.12966/oje.06.02.2014

Wang, Y., & Li, S. (2020). Issues, challenges, and future directions for multilingual assessment. Journal of Language Teaching and Research, 11(6), 914–919. https://doi.org/10.17507/jltr.1106.06