Developing Reliable Language Test Points as a Competency Testing Tool at High School Levels

Warsiman[1], Wandayani Goeyardi[2], Lilik Wahyuni[3]

{ warsiman@ub.ac.id[1], wanda_goey@ub.ac.id[2], lilikwahyuni@ub.ac.id[3]}

Universitas Brawijaya[123]

Abstract. Developing reliable test items is important for a teacher. In preparing the test items, one must pay attention to the level of difficulty and the level of distinguishing power, so that the results can distinguish smart students from less intelligent students. Test items that are very easy or very difficult, and test items that cannot distinguish between good and poor students, can provide invalid information on test results. The purpose of this study is to explain the test items that are trusted as competency test tools. This research method is descriptive with a qualitative approach. The results of the study explain that compiling reliable test questions is important to do. Therefore, in compiling test questions, it is necessary to pay attention to the level of difficulty and the level of distinguishing power, so that it can provide valid information, and it can place the test taker in the right place.

Keywords: Item test, reliable, competency test tool.

1 Introduction

Designing the ideal test item is far more important than simply ranking the test results obtained by students through unreliable test kits. Likewise, unreliable test items not only reduce the validity of the results but can further harm students. Professional teachers are teachers who do not only carry out routine learning tasks but also always carry out self-innovation including reflecting on their professional responsibilities in carrying out correct assessments [1].

This paper will explore reliable test items as a competency test tool in schools, from the point of view of the importance of a teacher setting test items that pay attention to the level of difficulty and level of differentiation, to obtain truly valid assessment results to determine the position of students in groups or groups. class. In addition, in this paper, the author would like to invite teachers in particular, and all parties with an interest in assessment, to conduct an analysis of test items based on the level of difficulty and level of differentiation before carrying out testing activities. This is done to obtain reliable test items and can be used to measure student competence. Not only that, reliable test items can be a vehicle for improving and perfecting learning activities [2]. Therefore, the good or bad results of the evaluation results depend on the correct measurement process [3].

ICEL 2022, November 07-08, Malang, Indonesia

DOI 10.4108/eai.7-11-2022.2329394

The results of the reliable test items are expected to determine students in the right position in the class. Test items that are arranged appropriately and appropriately can carry out their functions properly [4]. Meanwhile, the determination of test items as an unreliable assessment medium, the results are very detrimental to students. So far, wrong assessment activities often occur, and the impact is very hurtful to students, especially if the results of the assessment are used as a measure of passing or not passing, increasing or not increasing, and so on.

This phenomenon raises questions and at the same time doubts among the public about the professionalism of teachers. Against this professionalism, society demands that teachers equip themselves with adequate knowledge in their work as agents of change. He considered that so far the teacher had not shown himself as a professional in his work. This assumption is quite reasonable because so far the task of a teacher is considered a routine job that is equated with other jobs. We can see clearly how an unprofessional teacher is about these duties and responsibilities. As mentioned by Wijaya [5], teachers in their duties often do not care about the educational progress of their students. Children who get bad grades on a test, for example, continue the next day's lesson regardless of whether the lesson has been mastered or not.

Seeing the negative stigma, teachers or anyone with an interest in assessment should try to erase that image. Efforts that are deemed appropriate are to improve self-quality and performance. Without any change in self and performance, the negative stigma will always buzz around the world of education. As a result, public trust in teachers will decrease, and it can further damage the authority of the profession.

In line with the times, improvements in the field of quality education are getting louder and louder. Quality education is needed to fill the development of a nation. Today education is at the forefront. Education is considered the foundation of a country's development. If a country wants change and progress, it must pay attention to the larger education sector. Quality education is determined by how far the teacher's competence is. This is where the role of the teacher as an agent of change cannot be denied. Therefore, improving the quality of self for a teacher is a necessity. If these efforts are carried out continuously and seriously, the negative stigma of society toward teachers will be buried over time and the evidence shown.

Based on the explanation, the following problems can be formulated: 1) how to formulate the level of difficulty of the test items as a reliable student competency test tool, and 2) how to formulate the level of discriminating power of test items as a reliable student competency test tool.

2 Research methods

This study used the descriptive qualitative method. Therefore, researchers must have a clear picture of the aspects under study. The aspects studied in this study are the level of difficulty and the level of discriminating power in the preparation of test items. The data of this research are the learning outcomes of students in class VIII-E of SMP Negeri 6 Sidoarjo in the Indonesian language subject. The number of documents is 10 documents of student learning outcomes from 30 documents or several class VIII-E students of SMP Negeri 6 Sidoarjo. The document is taken randomly as a sample. Sampling is intended to facilitate researchers in making examples of how to calculate the index of difficulty level and level of test item power. The data document of student learning outcomes is processed to determine the level of difficulty and the level of differentiating power of the test items.

3 Results and discussion

3.1 Analysis of the Difficulty of Test Items

The level of difficulty of the test items is an important thing that must be known by the teacher or people with an interest in measurement/assessment. The level of difficulty is an indication of the quality of the overall test items administered. By knowing the level of difficulty, we can determine how difficult, moderate, easy, or too easy the test items we have compiled are, then we can determine whether or not the test items need to be revised, used, or discarded.

In simple terms to conclude the level of difficulty of the test items, we can observe through the scores obtained by students. If the average score of students who show most or all of them is high, then the test items are considered easy. On the other hand, the average score of students, which is mostly or completely low, indicates that the test items are difficult or very difficult.

A similar analysis can also be done by concluding how many test items can be answered correctly or answered incorrectly by the test taker. Test items that can be answered correctly by most or all students are considered easy or very easy test items. Conversely, if little or no one can answer, then the test item is difficult or very difficult.

To obtain more complete information about the difficulty level of a test, both the difficulty level of the entire test and each item, a teacher needs to conduct an analysis. The need for the level of difficulty to be analyzed is so that errors are not found in the preparation of test items that are very difficult or very easy, or in other words, the test items are sought so that they are not too difficult or not too easy so that it can be seen which participants have a high level of ability and which participants have a low level of ability.

Test items that are too easy or too difficult will provide invalid information on the test results. It is as if all students are smart for test items that can be answered easily, and as if all students are stupid for test items that are difficult for students to answer. Both possibilities are not following the normal circumstances that are commonly found in normal groups as is common in groups of students in most classes. As is known according to the normal curve principle and assumption, a normal group consists of group members who are generally normal and have normal level abilities except for a small number who have slightly above normal abilities and slightly below normal abilities. Because of this, the existence of tests or test items that have such extreme characteristics (too difficult or too easy) needs to be made aware by the teachers, because the questions do not appropriate expectations and assumptions of the normal curve.

The difficulty level is the ratio between the number of students' correct answers from the total number of students who took the test. More number of students who are able to answer the test items correctly, so the test items are easier, and the other way around.

To find out the level of difficulty of the test items, each skilled in the field of assessment sets each different of formula. However, those differences basically serve the same purpose. [6] provides the formula for finding the level of difficulty of the test items as follows below.

p = (JJB:JPT) x100%

Information:
p = difficulty level
JBB= number of correct answer
JPT = number of participants of test

For example, if the correct answer is only given by 15 out of 30 test takers, then the difficulty level can be expressed as 50% or 0.50. The obtained from a simple calculation as follows below.

p = (15:30) x 100%
= (0.5) x 100%
= 50%

As said by Witherington (Sudijono, 2006:371), that the difficulty index of the test items ranges from 0.00 to 1.00. That means, the lowest difficulty index is 0.00 and the highest is 1.00. The greatest difficulty index of 0.00 is an indication that the test items prepared by the teacher are in the too difficult category. As the other way around, if the difficulty index of the items is 1.00, it means that the item or test item is too easy.

How to provide an interpretation of the difficulty index of the item, Robert L. Thorndike and Elizabeth Hagen (1961, in [7])in their book Measurement and Evaluation in Psychology and Education stated as follows below:

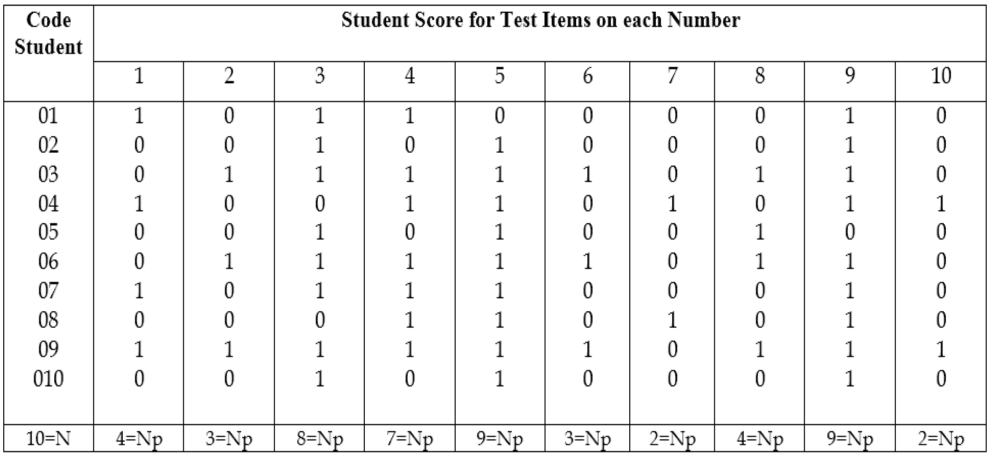
Proportion (P)	***Interpretation***
Less than 0.30	Difficult
0.30-0.70	Medium
More than 0.70	Easy

Meanwhile, [8] in a book entitled Psychological Education describes how to interpret the item difficulty index as follows below:

Proportion (P)	***Interpretation***
Less than 0.25	Difficult
0.25-0.75	Medium
More than 0.75	Easy

The technique for calculating the item difficulty index is presented in a random sample of ten documents of student learning outcomes of class VIII-E SMP Negeri 6 Sidoarjo who take the Indonesian language exam from 30 students. The questions are arranged in the form of an objective test by presenting 10 test items, and each test item can be answered correctly then it is given a weight of 1 and for each wrong answer a weight of 0 is given. After the exam is complete, corrections are made and a score is given. The index calculation refers to the calculation technique exemplified by [9] as follows.

Table 1. Student score for test items on each number

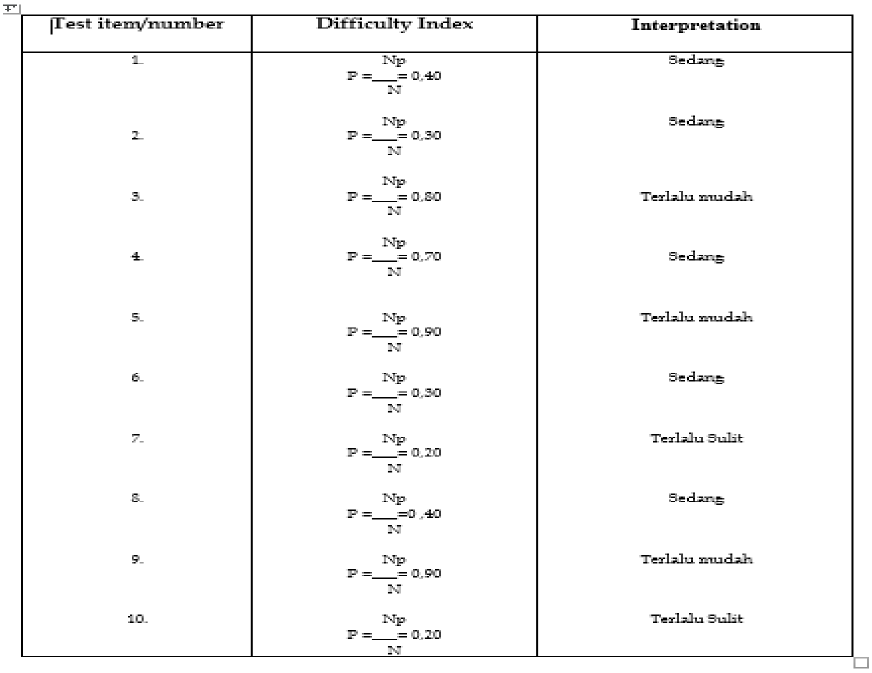


Code Student	Student Score for Test Items on each Number									
	1	2	3	4	5	6	7	8	9	10
01	1	0	1	1	0	0	0	0	1	0
02	0	0	1	0	1	0	0	0	1	0
03	0	1	1	1	1	1	0	1	1	0
04	1	0	0	1	1	0	1	0	1	1
05	0	0	1	0	1	0	0	1	0	0
06	0	1	1	1	1	1	0	1	1	0
07	1	0	1	1	1	0	0	0	1	0
08	0	0	0	1	1	0	1	0	1	0
09	1	1	1	1	1	1	0	1	1	1
010	0	0	1	0	1	0	0	0	1	0
10=N	4=Np	3=Np	8=Np	7=Np	9=Np	3=Np	2=Np	4=Np	9=Np	2=Np

Test item number 1 was answered correctly by students 01, 04, 07 and 09 (four students). For test item number 1, Np = 4, while N = 10. The difficulty index or P for test item number 1 is 4/10 = 0.40, so we can interpret that test item number 1 is in the medium category. For test items number 6 and 9 with Np of 3 and 9 respectively, then we can easily calculate the difficulty index numbers, respectively 3/10=0.30 (for item number 6) and 9/10=0.90 (for item number 9), so that the interpretation results that we give to test item number 6 are categorized as difficult test items, while test item number 9 is too easy test item category.

To observe the overall results of the analysis, consider the following table below:

Table 2. The difficulty level of the 10 test items



Test item/number	Difficulty Index	Interpretation
1.	$P = \frac{Np}{N} = 0,40$	Sedang
2.	$P = \frac{Np}{N} = 0,30$	Sedang
3.	$P = \frac{Np}{N} = 0,80$	Terlalu mudah
4.	$P = \frac{Np}{N} = 0,70$	Sedang
5.	$P = \frac{Np}{N} = 0,90$	Terlalu mudah
6.	$P = \frac{Np}{N} = 0,30$	Sedang
7.	$P = \frac{Np}{N} = 0,20$	Terlalu Sulit
8.	$P = \frac{Np}{N} =0 ,40$	Sedang
9.	$P = \frac{Np}{N} = 0,90$	Terlalu mudah
10.	$P = \frac{Np}{N} = 0,20$	Terlalu Sulit

3.2 Analysis of the differential level of test items

Beside analyzing the level of difficulty, the teacher also needs to do is analyze the level of differentiating power of the test items. The differentiating power of test items is how the test items can distinguish groups of students who are smart (upper group) and groups of students who are less (lower group) [10]. In other words, the test items can distinguish groups of students with high abilities and groups of students with low abilities. The higher the level of discriminating power of a test item, the higher the ability to be able to distinguish smart students from students who are less intelligent.

The level of distinguishing power as a characteristic of test items, extends from the lowest level to the highest, as does the level of difficulty of the test items. According to Djiwandono (1996:144) the range of the index referred to is as follows.

0.50 or more	:	considered good
antara 0.20 sampai 0.50	:	considered enough (medium)
less than 0.20	:	considered less
0	:	considered no descrimination
- (negative)	:	considered negative

As well as to the level of difficulty, the test items with the lowest level of discriminating power can not at all distinguish between groups of students who are smart and those who are less intelligent. Judging from the number of correct answers, the test items answered equally correctly by students from the smart group and the less intelligent group, for example, is one example of an invalid test item.

The level of discriminating power of test items will be higher than the lowest level if there is a difference between the number of correct answers produced by the two groups of intelligent and less intelligent. The greater the difference in the number of correct answers between the two groups, the greater the level of discriminating power of a test item. The highest level of discriminating power that can be achieved by a test item is 1.00, which means that all members of the smart group managed to give the correct answer, while no one from the less intelligent group managed to give the correct answer.

To find out determine the level of differentiating power of test items, each expert in the field of evaluation differs in formulating calculations. [6] for example, he gives the formula for finding the level of discriminatory power of test items as follows below.

$$D= (T-R):N$$

Information:
D = different level
T = *upper group*
R = *lower group*
N = number *upper group* and *lower group*

As example, if the number of members of group T and group R is set at 20 each, and among the answers to group T there are 14 correct answers and only 9 correct answers in group R, then:

$$\begin{aligned} D &= (14-9):20 \\ &= 5:20 \\ &= 0{,}25 \text{ or } 25\% \end{aligned}$$

Based on the explanation, the following problems can be formulated: 1) how to formulate the level of difficulty of the test items as a reliable student competency test tool, and 2) how to formulate the level of discriminating power of test items as a reliable student competency test tool.

Another way to find the level of discriminating power of test items can also follow the formula described by Hasan and Zainul (1991/1992:130-132) below.

$$D= \frac{B1-B2}{0.5T}$$

Information:

D = power difference;

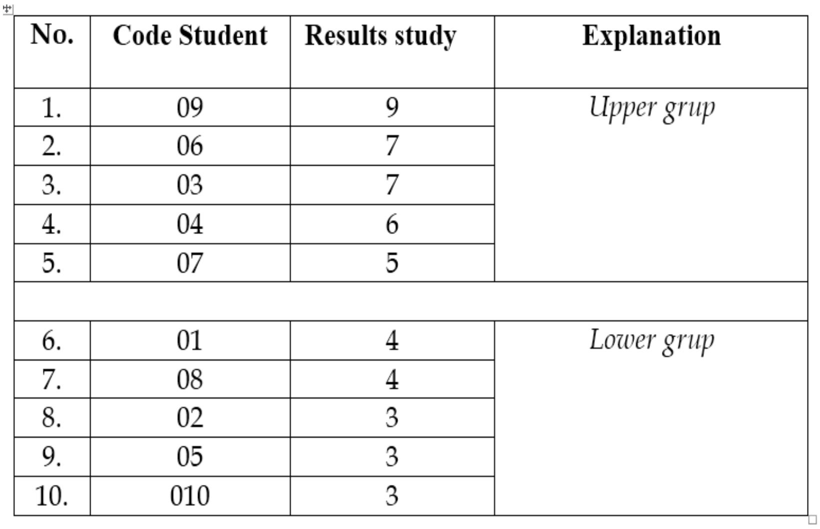
Ba = the number of upper group students who answered correctly for each question;

Bb = number of lower group students who answered correctly for each question;

T = the number of students (if the student is odd, then T = the number of students is less than one).

If this formula is applied to the sample of student learning outcomes of class VIII-E SMPN 3 Sidoarjo listed in table number 1 above, then to calculate the differentiating power of test item number 4, the following results will be obtained.

Table 3. The result study



No.	Code Student	Results study	Explanation
1.	09	9	*Upper grup*
2.	06	7	
3.	03	7	
4.	04	6	
5.	07	5	
6.	01	4	*Lower grup*
7.	08	4	
8.	02	3	
9.	05	3	
10.	010	3	

The top group who answered correctly test item number 4 were 5 students, while the bottom group had 3 students. The application of the formula is as follows:

$$D = \frac{Ba\text{-}Bb}{0.5T}$$

$$D = \frac{5-3}{0.5T}$$

$$D = \frac{2}{5}$$

$$D = 0.4$$

Based on these calculations, the test item number 4 has a difference of 0.4. The test items are categorized as medium level test items. To analyze other test items, you can follow the rules by applying the formula.

If we look for formulas for analyzing discrete power, there are many. Although different ways, but still have the same characteristics. For example, another way to find the level of discriminatory power of test items formulated by [10], the teacher can also use the following formula.

$$DP = \frac{U - T}{½\, T}$$

Information:

DP = discrimination index;

U = number of upper group students who answered correctly for each question;

L = the number of lower group students who answered correctly for each question;

T = number of opper group and lower group.

To use this formula, the teacher can choose which one is considered the easiest and most relevant to the problem, then apply it.

4 Conclusion

Every learning activity is always accompanied by an assessment. These two things become mandatory activities for a teacher to be able to determine the position of students in the class. The accuracy of an assessment depends on how well the teacher prepares the assessment tool. A professional teacher is required to be able to carry out assessment activities correctly. Errors in assessment activities not only reduce the validity of a result but can further hurt students because they will accept the consequences of the results obtained through unreliable assessment methods.

To create valid assessment results, teachers need to develop reliable test items. This means that the test items can distinguish which children are smart and which children are stupid. A test item that is very difficult, and impresses all stupid children, or conversely a very easy test item that impresses all smart children, is a form of an unreliable test item.

Correct and reliable assessment activities will only be carried out by a teacher who understands his duties and responsibilities. Therefore, efforts to always improve performance and professional work for teachers are a necessity. The demand for teachers to work professionally is a non-negotiable demand. Once he decides to work as a teacher, he should sincerely be the guardian of change.

For a teacher/educational practitioner or anyone with an interest in assessment, it is necessary to equip themselves with the ability to compose reliable test items. Reliable test items not only put students in their place but also give a sense of fairness.

References

[1] Riadi, Ahmad: Teacher competence in the implementation of learning evaluation. Ittihad Journal of Kopertais Region XI Kalimantan, 15(28), 52-67 (2017)

[2] Arifin, Zainal: Evaluation learning. Bandung: Rosdakarya Youth. (2011)

[3] Elviana: Analysis of items for evaluation of Islamic religious education learning using the anates program. Mudarrisuna Journal, 10(2), 58-74. (2020)

[4] Muluki, Ardilah, Patta Bundu, and Sukmawati: Quality analysis of odd semester test items for science subject class IV MI Radhiatul Adawiyah, Elementary School Scientific Journal, 4(1), 86-96 (2020)

[5] Wijaya, C. et al: Reform efforts in education and teaching. Bandung: Remadja Karya. (1988)

[6] Djiwandono, M. Soenardi: Language test in teaching. Bandung: Publisher ITB Bandung (1996)

[7] Sudijono, Anas: Introduction to educational evaluation. Jakarta: Raja Grafindo Persada (1996)

[8] Witherington: Educational psychology. (Translator M. Buchori). Bandung: Bapemsi family (1967)

[9] Arikunto, Suharsimi: Basics of educational evaluation. Jakarta: Earth Literacy (1992)

[10] Purwanto, Ngalim: Teaching evaluation principles and techniques. Bandung: Rosda Karya Publisher (2001)

[11] Brown, James Dean. (2002). Referenced Language Testing Criteria. Hawaii: Cambridge University Press.

[12] Hasan, S. Hamid and Asmawi Zainul. (1991/1992). Evaluation of Learning Outcomes. Jakarta: Ministry of Education and Culture. Director General of Higher Education.

[13] Lexy J. Moleong. (1990). Qualitative Research Methods. Bandung: Youth Rosda Karya.

[14] Muhajir, Noeng. (1998). Qualitative Research Methodology. Yogyakarta: rake Sarasin.

[15] Tuckman, Bruce W. (1975). Measuring Educational Outcomes: The Fundamentals of Testing. New York: Harcourt Brace Jovanovich.