Jurnal Penelitian dan Evaluasi Pendidikan
Volume 24, No 2, December 2020 (156-165)
Online: http://journal.uny.ac.id/index.php/jpep

CHARACTERISTICS OF MATH NATIONAL-STANDARDIZED SCHOOL EXAM TEST ITEMS IN JUNIOR HIGH SCHOOL: WHAT MUST BE CONSIDERED?

Atin Argianti[1*], Heri Retnawati[1]
[1]Department of Mathematics Education, Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Karangmalang, Depok, Sleman, Yogyakarta 55281, Indonesia
*Corresponding Author. E-mail: atinargianti.2018@student.uny.ac.id

ABSTRACT

This research aims to determine the characteristics of items for the Math National Standardized School Exam (NSSE) in Junior High School (JHS) in grade 9. This study is descriptive-explorative quantitative research. The samples chosen were 293 ninth-grade students' answers at state JHS of 3 Pati with a package of questions consisting of 30 items. The data collected is an NSSE test instrument and participants' answers at State JHS of 3 Pati 2018/2019, collected by documentation. Experts validated the NSSE instrument, and the characteristic items of the NSSE instrument were analyzed using the classical test theory approach using Quest program. The question items of the math NSSE test at state JHS of 3 Pati are generally moderately good. Based on the classical theory approach, the result of instrument validity from expert judgment was 0.924, while the validity of items was 17 (56.7%) of items were very valid. The reliability was 0.78 (reliable category). Generally, Math NSSE items are in the easy category with a percentage of 83.3%. The discrimination index results indicate that, in general, the NSSE items are in a moderate category with a percentage of 60% (18 items). The distraction effectiveness shows that NSSE items are in the functional category with a percentage of 50%.
Keywords: *item characteristics, mathematical test instrument, national-standardized school exam*

How to cite: Argianti, A., & Retnawati, H. (2020). Characteristics of Math national-standardized school exam test items in junior high school: What must be considered?. *Jurnal Penelitian dan Evaluasi Pendidikan, 24*(2), 156-165. doi:https://doi.org/10.21831/pep.v24i2.32547

INTRODUCTION

Assessment is an important thing in education to identify whether the education is successful or not (Retnawati, 2016). As Previously mentioned, assessment is the process of gathering and processing information to measure the achievement of student learning outcomes (Regulation of the Minister of Education and Culture of the Republic of Indonesia No. 4 of 2018). In addition, the purpose of the assessment is (1) Assessment of learning outcomes by educators aims to monitor and evaluate the learning process, learning progress, and continuous improvement of student learning outcomes. (2) Assessment of learning outcomes by the education unit aims to assess the achievement of Graduates' Competency Standards for all subjects. (3) Assessment of learning outcomes by the government aims to assess the achievement of graduate competencies nationally on certain subjects. Assessment is carried out by learning, namely the process of interaction between students, between students, educators and learning resources in a learning environment (Regulation of the Minister of Education of Culture of the Republic of Indonesia No. 23 of 2016).

Secondary school student learning outcomes can be measured through assessment of learning outcomes by educational units, one of which is a school exam. School/madrasah examination are activities carried out to measure the achievement of student competencies as

ISSN: 2685-7111 (print)
2338-6061 (online)

doi:https://doi.org/10.21831/pep.v24i2.32547

recognition of learning achievements and/or completion of the educational unit (Regulation of the Minister of Education of Culture of the Republic of Indonesia No. 23 of 2016). By the assessment, the teacher can evaluate learning including in carrying out assessment of learning outcomes and instruments for assessing students' abilities (Arifin, 2012). As previously mentioned, teacher has task to evaluating quality of learning through the result of assessment (Retnawati et al., 2017). It has a goal to measure the success of learning followed by students in mastering the competencies that have been determined.

One of the measures that can be measured is the assessment of knowledge, for example mathematics. The education unit evaluates mathematics, one of which is *NSSE*. *NSSE* is an activity to measure student competency performance carried out by the education unit by referring to Graduates' Competency Standards to gain recognition of learning achievement (Regulation of the Minister of Education and Culture of the Republic of Indonesia No. 4 of 2018). In addition, the value of *NSSE* needs to be reported by the education unit for improving and equitable distribution of education quality (Regulation of the Minister of Education and Culture of the Republic of Indonesia No. 4 of 2018). Therefore, special attention is needed to *NSSE*. The activity is carried out at elementary education, junior high education, and senior high education levels by certain lessons, one of which is mathematics.

Test instrument is used for identifying mathematical knowledge. It is supported by Popham (2009) who states that the test instrument given to the test participants will show how the achievement or ability measured. A quality test will show how the actual test results are achieved so that a quality measuring instrument is needed (Kartowagiran et al., 2018). Analysis of item characteristics can produce quality tests (Gronlund, 1998; Retnawati, 2016).

The quality of the test device can be identified by conducting qualitative and quantitative analysis. Qualitative test analysis can be conducted by examining the suitability of the items based on the basic abilities and indicators to be measured and whether tests items have met the requirements of the material, construction, and language aspects. Meanwhile, item analysis by quantitative method can be conducted using two approaches, namely classical test theory and item response theory. However, this study uses classic test theory approach because the data obtained are the result of student's work after working *NSSE*.

Classical test theory is widely used because it does not require large respondents (more than 100) and is easy to apply, so measurement by *NSSE* in educational units does not involve many respondents. Quantitative analysis according to the classical test theory approach produces item characteristics which include the index of difficulty (p), differentiation (d), and effectiveness of distractors. In addition, reliability of test questions, and standard measurement errors can also be identified by the classical theory approach of quantitative research.

According to Ebel and Frisbie (1991), a test is said to be good if it has fulfilled the characteristics of the test, namely a test that has validity, relevance, balance, efficiency, specificity, difficulty, and reliability. Related to this matter, research about the characteristics of the *NSSE* test items is needed.

RESEARCH METHOD

The research was a descriptive-explorative quantitative research. The population of this study were all state junior high schools in Pati City. The selection of this research samples was purposive sampling because it has the purpose of knowing the character of *NSSE* instrument. Thus, the samples chosen are 293 ninth-grade students at state JHS of 3 Pati. The school was in the center of city Pati. The average age of students who took the test was 13 years. The students had medium to high ability.

The data used in the study were test instrument that included question items, and student's answer sheets obtained from math *NSSE* at Pati District Middle School 2018/2019. The question items were made by Subject Teacher Forum at Pati Regency. The data were sec-

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

doi:https://doi.org/10.21831/pep.v24i2.32547

ondary data collected by documentation. The type of math *NSSE* questions used was multiple choice with a package of questions consisting of 30 items. After the data were collected, the researchers conducted data calculation to find out the validity. From the collected data, the researchers also analysed the reliability, difficulty index, discrimination index, and distractor effectiveness of math *NSSE* test instrument. The data analysis technique in this study used classical test theory approach by using Quest program. Each item analyzed was characterized based on classical test criteria.

Validity

Validity was used to prove the degree of facts and theories that support the interpretation of test scores. Content validity was proven by using expert judgment of three experts. To find out the results of expert judgment were calculated by using the Aiken index. The three experts validated the items in the test instrument by noticing the suitability of the items with the competency standard of graduates by using a blueprint. The validity criteria used in the research based on Retnawati (2016). It can be seen in Table 1.

Table 1. Criteria for the Validity of Items in the Math NSSE

Criteria	Number of Items
Low	$v \leq 0.4$
Moderate	$0.4 < v \leq 0.8$
Very Valid	$v > 0.8$

Reliability

Reliability was seen from the reliability coefficients. Reliability coefficients of 0.8 and up are typically regarded as moderate to high, while coefficients below 0.6 are low (Andrade & Heritage, 2018).

Difficulty Index

Difficulty index is usually expressed in a proportion between 0.00 and 1.00. The formula of difficulty index item is as presented in Formula (1), where P_i = index of difficulty item to i, i = number of item, n = number of students who answered the item correctly, and N = the number of students who answered the item. Meanwhile, the difficulty level criteria based on (Retnawati, 2016) are presented in Table 2.

$$P_i = \frac{n}{n} \text{ ..... (1)}$$

Table 2. Classification of the level difficulty

Coefficient of Level	Criteria
0.00 < DI ≤ 0.30	Difficulty
0.30 < DI ≤ 0.70	Moderate
0.70 < DI ≤ 1.00	Easy

Discrimination Index

Discrimination index is calculated by index of discrimination is the biserial point. According to Crocker and Algina (1986), biserial point coefficient is determined by Formula (2), where ρ_{pbis} = biserial point correlation, μ_+ = average score of test participants who answer correctly the item, μ_τ = average total score, σ_τ = standard deviation total score, p = proportion the number of participants who answered correctly, and $q = 1 - p$. In addition, the discrimination criteria are based on Arikunto (2012), as presented in Table 3.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

doi:https://doi.org/10.21831/pep.v24i2.32547

$$\rho_{pbis} = \frac{\mu_+ - \mu_\tau}{\sigma_\tau}\sqrt{\frac{p}{q}} \quad \text{..... (2)}$$

Table 3. Classification of Discrimination Index

Coefficient	Criteria
$0.00 < \rho_{pbis} \leq 0.20$	Poor
$0.20 < \rho_{pbis} \leq 0.40$	Enough
$0.40 < \rho_{pbis} \leq 0.70$	Good
$0.70 < \rho_{pbis} \leq 1.00$	Very Good

Distractor Effectiveness

Distractor effectiveness could be said to function well at least it was chosen by 5% of the test participants (Arikunto, 2012). Thus, distractors chosen by less than this percentage are considered less effective.

FINDINGS AND DISCUSSION

Findings

The analysis results are to determine content validity, reliability, difficulty index, discrimination index, and distractor effectiveness. The analysis results of the items are compared with the criteria to determine whether the items are received, revised, or rejected based on the difficulty index, discrimination index, and distractor effectiveness.

Validity

The validity of the math *NSSE* instrument can be seen in Table 4. It can be seen that there are 13 items having moderate validity or 43.3% of items are moderately valid. Meanwhile, there are 17 items having high validity or 56.7% of items are very valid. Furthermore, the results of calculations was obtained that the expert agreement index for content validity was 0.924. It can be interpreted that the instrument are very valid.

Table 4. The Analysis Results of Validity of Items

Criteria	Number of Items	Item Numbers
Moderate	13	1, 4, 6, 7, 8, 11, 14, 15, 18, 19, 20, 21, 30
Very Valid	17	2, 3, 5, 9, 10, 12, 13, 16, 17, 22, 23, 24, 25, 26, 27, 28, 29

Reliability

Based on the analysis results by using Quest, it was obtained that the *Internal Consistency* value was 0.78. It means that the test instrument is reliable.

Index of Difficulty

The difficulty index was calculated by using Quest program. The difficulty index of math *NSSE* items can be seen in Table 5. It can be concluded that there are 25 items having easy category or 83.3% of items are easy. Besides, there are five items having moderate category or 16.7% of items are moderate, but there is no item having difficult category.

Table 5. The Analysis Results of Difficulty Index of Items

Criteria	Item Numbers
Easy	1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15, 16, 17, 19, 20, 21, 22, 23, 24, 26, 28, 29, 30.
Medium	9, 14, 18, 25, 27.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

doi:https://doi.org/10.21831/pep.v24i2.32547

Discrimination Index

The results of the analysis with the Quest program obtained the data presented in Table 6. It can be concluded that: there are 18 items having enough discrimination index or 60% of items are enough, there are 11 items having good discrimination index or 36.7% of items are good, and there is one item having poor discrimination index or 3.3% of items are poor. Meanwhile, there is no item having very good discrimination index.

Table 6. The Analysis Results of Discrimination Index of Items

Criteria	Item Numbers
Poor	13
Enough	1, 2, 3, 4, 5, 6, 8, 10, 11, 15, 16, 17, 20, 22, 24, 26, 28, 29
Good	7, 9, 12, 14, 18, 19, 21, 23, 25, 27, 30

Distractor Effectiveness

The distractor effectiveness was calculated by using the Quest program. The distractor effectiveness of math *NSSE* items can be seen in Table 7.

Table 7. The Analysis Results of Distractor Effectiveness of Items

Percent	> 5%	< 5%
Item numbers	1, 7, 8, 9, 11, 12, 14, 16, 17, 18, 23, 24, 25, 27, 30	2, 3, 4, 5, 6, 10, 13, 15, 19, 20, 21, 22, 26, 28, 29

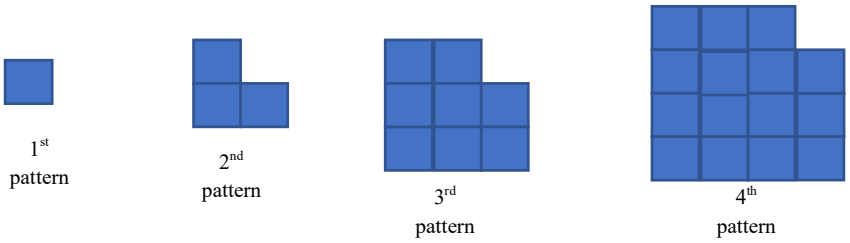
Based on Table 7, it can be concluded that there are 15 items having distractor effectiveness with more than 5% chosen by participants and 15 items having distractor effectiveness with less than 5% chosen by participants.

Analysis Results about Question Items

One of the easy questions is the 5th number. It is proved by the fact that 246 of 293 students correctly answer. Item the difficulty index is 0.959. It means that the proportion of student who correctly answer is 95.6%. The discrimination index of this item is 0.24.

Item Number 5

Look at the pattern below.



How many unit squares are in 6th pattern?
35
74
82
85

Solution:
1st pattern = 1
2nd pattern = 1.2 + 1 = 3
3rd pattern = 2.3 + 2 = 8
4th pattern = 3.4 + 3 = 15
5th pattern = 4.5 + 4 = 24
6th pattern = 5.6 + 5 = 35

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

doi:https://doi.org/10.21831/pep.v24i2.32547

The questions are easy for students to understand because the information in the question is clear and the question are still classified as routine. The students only need the skills to apply formulas to solve the question. Meanwhile, students' abilities are already high. The question must be increased cognitive level such as adding information to deceive.

Item Number 9

One of the moderate questions is the 9th number. It is proved by the fact that 134 of 293 students correctly answer. Item the difficulty index is 0.577. It means that the proportion of student who correctly answer is 57.7%. The discrimination index of this item is 0.47.

It is given:
$S = \{natural\ numbers\ less\ than\ 10\}$
$P = \{odd\ number\ \ less\ than\ 10\}$
$Q = \{factor\ number\ of\ 8\}$

The set of complement $(P \cup Q)$ is ...
{1,2,3,4,5,7,8,9,10 }
{1,2,3,4,5,7,8,9}
{6,10}
{6}

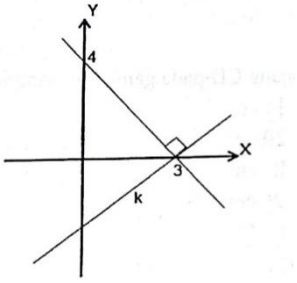
Solution:
$S = \{1,2,3,4,5,6,7,8,9\}$
$P = \{1,3,5,7,9\}$
$Q = \{1,2,4,8\}$
$(P \cup Q) = \{1,2,3,4,5,7,8,9\}$
$(P \cup Q)^c = \{6\}$

The problem can deceive students because in the stem of the question, students may find it difficult to distinguish between the "union" symbol and the "intersection" symbol. Besides, there is symbols written in word namely "complement". The good question should be written "The set of $(P \cup Q)^c$".

Item Number 14

The second question of easy questions is the 14th number. It is proved by the fact that 171 of 293 students correctly answer. Item the difficulty index is 0.7. It means that the proportion of student who correctly answer is 70%. The discrimination index of this item is 0.52.

Notice the picture below!



The equation of k line is ...
$4x + 3y - 12 = 0$
$4x + 3y + 12 = 0$
$3x - 4y - 9 = 0$
$3x - 4y + 9 = 0$

Solution:
The gradient of points (0.4) and (3.0) is given by $m_1 = \frac{4-0}{0-3} = -\frac{4}{3}$
Since k line is perpendicular to points (0.4) and (3.0), the gradient m_2 is given by
$m_2 = -\frac{1}{m_1} = -\left(\frac{1}{-\frac{4}{3}}\right) = \frac{3}{4}$
k line is passed through the point (3.0), so the coordinates of k line is (3.0)
Using the result $y - y_0 = m\,(x - x_0)$ gives $y - 0 = \frac{3}{4}(x - 3)$

$$4y = 3x - 9$$
$$3x - 4y - 9 = 0$$

Therefore, the equation of k line is $3x - 4y - 9 = 0$.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

doi:https://doi.org/10.21831/pep.v24i2.32547

The questions are too easy for students to understand because the information in the question is clear and the question are still classified as routine. The students only need the skills to apply formulas to solve the question. Besides, understanding the concept of algebraic operations will affect the student work process (Shantika & Istiyono, 2019). The question can be turned into word problems without displaying pictures on the questions and the level of questions can be increased. The distractor effectiveness is improved again.

Discussion

According to Retnawati (2017), an instrument is considered to have a good quality based on the validity and reliability as well as the characteristic of the instrument component. Based on the results obtained from Table 4, it can be seen that 13 items or 43.3% of items are moderately valid, while 17 items or 56.7% of items are very valid. It is supported by the content validity index which has value of 0.924. It can be interpreted that the test instrument is very valid for all items. In addition, validity is a tool to consider all important things in the development of a test. Other experts suggest that the validity of a measuring instrument can measure what should be measured (Allen & Yen, 1979). Linn and Gronlund (1995) explain that validity refers to the adequacy and feasibility of interpretations made from assessments, with regard to specific uses.

Accordingly, the test instrument can be said as the high-quality instrument. According to Linn and Gronlund (1995), the high-quality instrument that is good (correct, valid, and appropriate). Validity also can measure the success of students in the learning process in a certain period (Nitko, 1996). Meanwhile, the moderately valid items can be revised by considering the relation between questions and question indicators.

According to the analysis result by using Quest program, the *Internal Consistency* value was 0.78. It can be said that the test device is very reliable. The higher the correlation questions, the higher the reliability (Nunally, 1978). In addition, the reliability coefficient is closely related to the *standard error of measurement* (SEM)/measurement error, based on the calculation, it is known that the standard deviation is 3.49 so that the SEM value of 1.637 is obtained. SEM is an estimate of the number of errors in the test score. The smaller the SEM, the higher realibility of a test device. It can be said that the test instrument has a high accuracy level.

Based on Table 5, there are 25 items having easy category or 83.3% of items are easy, there are five items having moderate category or 16.7% of items are moderate, but there is no item having difficult category. It indicates that the proportion of the math *NSSE* items is not balanced. Ideally, according to Arifin (2012), a good proportion of difficulty index should be spread normally. Meanwhile, Arikunto (2012) suggested that a good question is one that is not too easy and not too difficult. The question that is too easy does not stimulate students' hard effort to solve it. Conversely, the question that is too difficult will cause students to become discouraged and not to be enthusiastic in solving it because it is beyond their ability. In addition, the difficulty index of question items can be identified from the student's ability. The higher the student's ability, the lower the difficulty index. It means that the question items have a low difficulty index if students can answer easyly and correctly. Conversely, the question items have a high difficulty index if students have difficulty answering that questions.

Based on Table 6, there are 18 items having enough discrimination index or 60% of items are enough, there are 11 items having good discrimination index or 36.7% of items are good, and there is one item having poor discrimination index or 3.3% of items are poor. Meanwhile, there is no item having very good discrimination index. It can be said that the math *NSSE* items has moderately good quality based on the discrimination index. A good question item is an item that has a discrimination index more than 0.2 as stated by Fernandes (1984). According to Ebel and Frisbie (1972), an item is said to be quality if the discrimination index is at least 0.41.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

doi:https://doi.org/10.21831/pep.v24i2.32547

The results obtained from the difficulty and discrimination index of the math NSSE items need to be considered because according to Crocker (2008), one of the causes of poor discrimination is the index of difficulty of the question. Less and good enough questions need to be reviewed both in terms of the difficulty and the discrimination index. Discrimination index of a problem is the ability of a question to distinguish high-ability students from low-ability students. Logically, smart students will certainly be able to answer than low-ability students. Also, students do not believe in themselves or guess in answering the questions due to the lack of students' understanding of the concept (Istiyani et al., 2018).

Based on Table 7, there are 15 items having distractor effectiveness with more than 5% chosen by participants and 15 items having distractor effectiveness with less than 5% chosen by participants. Distractors do not function properly will make the item easier. Besides, question items cannot distinguish students who master the material being asked and students who do not master the material being asked.

The quality of this math *NSSE* instrument can be seen from its validity because there are still valid questions. The quality of the instrument can also be seen from its reliability. In this case, the math *NSSE* instrument is reliable. Furthermore, it can be seen from the index of difficulty that it must have a balanced proportion in order to get good learning achievement. In addition, the quality of the question instrument can be seen from the function of the distractor so that there can be a tendency to attract student to choose it. In addition, there are 50% of items that do not have a good distractor, so that most students choose the right answer. According to Brawn (Fernandes, 1984), distractors were said to be good at least it is chosen by 2% of all participants. Meanwhile, Nitko (1996) said that a distractor functions if at least it is selected by a test participant from a low group.

In compiling multiple choice questions, the ability to arrange alternative answers is a very important aspect. The use of a distractor that is not good will reduce the quality of the question. The results of research conducted by Attali and Bar-Hillel (2003) concluded that both test takers and question makers have the same tendency to choose answers or the place of answer key. This, absolutely, increases the chance of the test taker to guess the answer. The high ability of test taker to guess will decrease the discrimination index. Meanwhile, the low discrimination index will give the homoguneous scores. The more homogeneous scores are obtained, the weaker the reliability of the question will be (Allen & Yen, 1979). Thus, the question that is invalid and has low reliability can also be caused by the students' ability to guess in answering the tests given.

Beside the difficulty and discrimination index, in making decisions, it is also necessary to pay attention to the abilities of the participants/respondents, because basically, here is one of the disadvantages of this classic theory, namely the interrelationship between the characteristics of the items and participants. This is based on Naga (1992) that in the classical theory, the characteristics of items always depend on the group of participants.

The quality of *NSSE* math instrument can be seen from the level of the item's difficult. Meanwhile, the difficulty level of the items is based on the student's ability to understand the material taught by the teacher. According to Fernandes (1984), question item that results a mean score of around 50% of the maximum score can be said that the item has good index of difficulty. Meanwhile, Thomas and Dawson (1972) explained that items that had a difficulty level of 0.25 - 0.75 were said to be good.

NSSE is one of the determinants of graduation aside from the national-standardized examination (*NSE). NSE* functions as one of the considerations for quality mapping education units, the basis for selection to enter the next level of education, determining the graduation of students from programs and/or educational units, basic guidance, and assistance to educational units to improve the quality of education (Sunarti & Anggraini, 2013). Certainly, both of them have differences in terms of scope, level of difficulty, and management. *NSSE's*

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

doi:https://doi.org/10.21831/pep.v24i2.32547

scope is in accordance with regional provisions, but still in accordance with national examination standards, while the *NSE* scope is carried out nationally. The level of difficulty is different and remains in accordance with national examination standards. In management, *NSE* is managed by the government while *NSSE* is managed by the teacher or *subject teacher forum*.

CONCLUSION

Based on the discussion of the results obtained, it can be concluded that the question items of math *NSSE* test at state JHS of 3 Pati is generally moderately good. Based on the classical theory approach, the test instrument has validity value of 0.924 which means that it is very valid for all items. There are 13 items or 43.3% of items are moderately valid, while 17 items or 56.7% of items are very valid. The internal concistency value is 0.78 which means the test instrument is reliable. The difficulty index of the question items is included in the easy category. There are 25 items having easy category or 83.3% of items are easy, there are five items having moderate category or 16.7% of items are moderate, but there is no item having difficult category. The discrimination index of the items has moderately good category. There are 18 items having enough discrimination index or 60% of items are enough, there are 11 items having good discrimination index or 36.7% of items are good, and there is one item having poor discrimination index or 3.3% of items are poor. The distractors of the items generally function well. There are 15 items having distractor effectiveness with more than 5% chosen by participants and 15 items having distractor effectiveness with less than 5% chosen by participants.

Test instrument is used for identifying mathematical knowledge. Thus, for teachers and prospective teachers, it is important to know the rules of quality question writing such as validity, reliability, index of difficulty, discrimination index, and effectiveness of distractor. Besides, the teacher and all parties involved in making exam questions must consider the proportion of questions to be balanced between the questions that are, easy, medium, and difficult because each school has different student characteristics. For stakeholders of State JHS of 3 Pati, it is suggested to notice many aspects namely the readiness of students and the test instrument that will be used before the implementation of *NSSE*. This study can also be used as an evaluation material for learning at State JHS of 3 Pati. For the next researchers, it can be a reference in conducting the similar research or analyzing other subjects.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Brooks/Cole.

Andrade, H. L., & Heritage, M. (2018). *Using formative assessment to enhance learning, achievement, and academic self-regulation* (1st ed.). Taylor & Francis.

Arifin, Z. (2012). *Evaluasi pembelajaran*. Remaja Rosdakarya.

Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan* (2nd ed.). Bumi Aksara.

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109–128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

Crocker, L. (2008). *Introduction to classical and modern test theory*. Holt, Rinehard and Wiston.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehard and Winston.

Ebel, R. L., & Frisbie, D. A. (1972). *Essentials of educational measurement* (3rd ed.). Prentice Hall.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061

doi:https://doi.org/10.21831/pep.v24i2.32547

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice-Hall of India.

Fernandes, H. J. X. (1984). *Evaluation of educational programs*. National Education Planning Evaluation and Curriculum Development.

Gronlund, N. E. (1998). *Assessment of student achievement* (6th ed.). Allyn and Bacon.

Istiyani, R., Muchyidin, A., & Rahardjo, H. (2018). Analisis miskonsepsi siswa pada konsep geometri menggunakan Three-Tier Diagnostic Test. *Cakrawala Pendidikan*, *37*(2), 223–236. https://doi.org/https://doi.org/10.21831/cp.v37i2.14493

Kartowagiran, B., Munadi, S., Retnawati, H., & Apino, E. (2018). The equating of battery test packages of mathematics national examination 2013-2016. *SHS Web of Conferences*, *42*(00022), 1–6. https://doi.org/10.1051/shsconf/20184200022

Linn, R. L., & Gronlund, N. E. (1995). *Meansurement and assessment in teaching* (7th ed.). Prentice-Hall.

Naga, D. S. (1992). *Pengantar teori sekor pada pengukuran pendidikan*. Besbats.

Nitko, A. J. (1996). *Educational assessment of students* (2nd ed.). Merrill an imprint of Prentice Hall.

Nunally, J. C. (1978). *Psychometric theory* (2nd ed.). McGrawHill.

Popham, W. J. (2009). *Instruction that measures up: Successful teaching in the age of accountability*. ASCD.

Regulation of the Minister of Education and Culture of the Republic of Indonesia No. 4 of 2018 concerning the Learning Outcome Assessment by Educational Units and the Government. (2018).

Regulation of the Minister of Education of Culture of the Republic of Indonesia No. 23 of 2016 concerning the Educational Assessment Standard. (2016).

Retnawati, H. (2016). *Analisis kuantitatif instrumen penelitian*. Parama.

Retnawati, H. (2017). Learning trajectory of item response theory course using multiple softwares. *Olympiads in Informatics*, *11*, 123–142. https://doi.org/10.15388/ioi.2017.10

Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why are the mathematics national examination items difficult and what is teachers' strategy to overcome it? *International Journal of Instruction*, *10*(3), 257–276. https://doi.org/10.12973/iji.2017.10317a

Shantika, E. G., & Istiyono, E. (2019). A diagnosis of students' errors in answering the mathematics test in senior high school. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *23*(2), 129–143. https://doi.org/10.21831/pep.v23i2.16370

Sunarti, S. & Anggraini, D. (2013). Pengembangan bank soal dan pembahasan ujian nasional berbasis multimedia pembelajaran interaktif dengan macromedia authorware 7.0. *Jurnal Cakrawala Pendidikan*, *31*(3), 394–408. https://doi.org/10.21831/cp.v0i3.1138

Thomas, G., & Dawson, J. B. (1972). *Item analysis and examination statics*. The Union of Educational Institutions.

ISSN (print) 2685-7111 | ISSN (online) 2338-6061