How good are the characteristics of the mathematics national standardized school examination items in vocational school?

A Z Fathin and H Retnawati
Department of Mathematics Education, Graduate Program, Universitas Negeri Yogyakarta, Indonesia
E-mail: aizzazakkiyatul.2018@student.uny.ac.id

Abstract. National Standardized School Examination is one of the exams used to determine the graduation of vocational school students. This study aims to determine the characteristics of mathematics National Standardized School Examination items in 12th Vocational School. This research is a descriptive explorative quantitative research. The data used is the mathematics National Standardized School Examination kits at a vocational school in the city of Yogyakarta and participant’s answers collected with documentation. These assessment kits are validated by experts and analyzed its characteristics using the classical test theory approach. The results of the study indicate that content validity obtains the Aiken index value (V) of 0.94, which means that this instrument is very valid for all items. Reliability for this instrument is categorized as a medium with a reliability coefficient of 0.3623. The results of the characteristic analysis showed that 18 items are having moderate difficulty, and there are 17 difficult questions. Based on the discrimination index, there are 6 items included in the good category, and the other 19 items are categorized as not good. Furthermore, in terms of the effectiveness of the distractors, this instrument can be said to be effective because 82.86% of the items have effective distractors.

1. Introduction
Assessment is an important matter in education. According to the Indonesian Ministry of Education and Culture Regulation (Permendikbud) No. 23, the Year 2016, about educational assessment standards [1], assessment is the process of gathering and processing information to measure student learning outcomes. Indonesian Ministry of Education and Culture Regulation (Permendikbud) No. 23, the Year 2016, [1] on article 4 states that the assessment of learning outcomes aims to monitor and evaluate the process and progress of learning, continuous improvement of student learning outcomes and to assess the achievement of student competency standards. While according to Ebel and Frisbie [2] in education, the main function of assessment is to measure the students' achievements and to motivate students' learning.

One of the assessment forms that can be done is by giving a test to students. Tests are tools used to assess abilities, skills, and expertise related to education and the world of work to place individuals efficiently and objectively in various educational positions and organizations [3]. A quality test is needed for obtaining the result of the assessment which really describes the achievements of the test participants [4]. A quality test consists of good test items. The test items developed should be able to measure students’ abilities accurately and be able to distinguish between low and high ability students, and ensure that the items function according to their usage [5].

Good quality tests must consist of good items [6]. So, the analysis of items test is needed. There are two types of item analysis; qualitative and quantitative analysis. Qualitative analysis is called the review of items through three aspects, namely the substance, construction, and language aspects [7]. One of the
ways in qualitative analysis is to prove content validity, while the quantitative analysis is to know the quality of items test empirically [8]. Tests that have been developed well must be tested to a number of samples, and then the responses were used to analyze the characteristics of the items [9]. Analysis of the characteristics of items can use the classical test theory and item response theory. This study used the classical test theory. From the analysis of the characteristics using classical test theory, it will be known reliability, level of difficulty, item discrimination, and the effectiveness of distractors.

In Indonesia education, one of the tests in the assessment of learning outcomes is the National Standardized School Examination. According to Indonesian Ministry of Education and Culture Regulation (Permendikbud) No. 4, the Year 2018, about the assessment of learning outcomes by the education level and the government [10], National Standardized School Examination is an activity to measure student competency achievement carried out by the Education Level by referring to graduates’ competency standards to obtain recognition for learning achievement. National Standardized School Examination is one of the graduation requirements of an education unit. This is in accordance with the Indonesian Ministry of Education and Culture Regulation (Permendikbud) No. 4, the Year 2018, [10] on article 1, which states that students are declared to have graduated from education level/program after passing the level/program examination. In other words, the result of this test determines the graduation of the student. Therefore, the National Standardized School Examination test is expected to be a good quality instrument that can provide valid and objective information. Exam questions that have poor quality can provide information that is not in accordance with student achievement. This can be detrimental to students because they have provided information that is incorrect and misleading in making student graduation decisions.

Related to the explanation, this research is needed on the item characteristics of the National Standardized School Examination in a vocational school. This result research is expected to be able to obtain test instruments of good quality for the purpose of developing further tests.

2. Method
This research is a descriptive explorative quantitative research that aims to describe the characteristics of mathematics National Standardized School Examination items test in grade 12th Vocational School. Data collection was carried out through documentation of the mathematics National Standardized School Examination kits in a vocational school in Yogyakarta, Indonesia, which consisted of blueprint and mathematics National Standardized School Examination questions. These questions consist of 35 multiple choice questions with 5 answer choices where an option is the answer key, and the other is the distractor. Then data collection was also carried out through documentation of the responses from 286 students.

The data analysis technique used in the research was qualitative (theoretical) and quantitative (empirical) analysis. Qualitative analysis was done by proving the validity of the instrument. The validity of an instrument is the extent to which the measuring instrument is able to measure what should be measured. The validation carried out in this research was content validity by considering expert judgments. The content validity is carried out to find out the extent to which the items in the instrument represent the components in the entire area of the object to be measured and the extent to which the items reflect the behavioral characteristics to be measured [6]. Based on the responses’ expert judgments, the expert agreement index was calculated using the formula Aiken (V). Quantitative analysis was carried out to find out the characteristics of these test items in terms of reliability, the level of difficulty, item discrimination, and the effectiveness of the distractor. Characteristic analysis in this study was carried out based on classical test theory.

3. Findings

3.1. Content Validity
The content validity in this study used the agreement of 5 experts. Based on the result of it, the Aiken index (V) was used to determine the agreement of the rater (expert) on the suitability of the items with
the indicators to be measured using the item. From the results of the V Aiken index calculation, the items were categorized. If the index is less than or equal to 0.4, it is said that the validity is low, 0.4-0.8 is said to be moderate validity, and if it is more than 0.8, it is said to be very valid [6]. The content validity of the mathematics National Standardized School Examination items test is presented in Table 1.

| Criteria          | Item test            | Aiken index |
|-------------------|----------------------|-------------|
| Low               | 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35 | 0.94 |
| Moderate          | 10                   |             |

Table 1 shows that 34 test items were categorized as very valid and 1 test item was categorized as medium. Furthermore, the V Aiken index on the overall items is 0.94. It means that this instrument is very valid for all items. Thus, it means that the validators agreed that 94% of the total number of items have been relevant to the standard of graduation competence, basic competence, indicators, and cognitive levels.

3.2. Reliability
The reliability of a test is a coefficient that shows the degree of permanence or consistency of the result of the tests. The reliability of a test ($\rho$) is generally expressed numerically in the form of a magnitude coefficient between -1 to 1 [6]. Based on the statistical analysis for tests obtained by calculating the alpha coefficient using the Cronbach alpha formula, which is the reliability coefficient. Then the test items can be categorized based on the reliability coefficient obtained. If the coefficient is at 0-0.2 the reliability is very low, 0.2-0.4 is said to be moderate reliability, 0.4-0.6 reliability is enough, 0.6-0.8 reliability is high, and 0.8-1 reliability is very high [11]. The reliability coefficient of the mathematics National Standardized School Examination test is 0.362. It is categorized as medium.

3.3. Characteristic of Test Items
Analysis of the items characteristics in this study used the classical test theory. The result of this analysis can show the criteria that determine the quality of the test based on the level of difficulty, item discrimination, and the effectiveness of the distractors. The difficulty level of the items in classical test theory is defined as the proportion of examinees who answer the item correctly. According to Allen and Yen [12], items with a good degree of difficulty is the interval of 0.3 – 0.7. The item test is considered difficult if the difficulty level is below 0.3. And item with the difficulty level above 0.7 indicates the easy item. The result of the difficulty level analysis of mathematics National Standardized School Examination tests is presented in Table 2.

| No. | Material | Criteria                  |
|-----|----------|---------------------------|
|     |          | Moderate item (0.3 – 0.7) | Difficult item (> 0.7) |
| 1.  | Algebra  | 8 items (1, 2, 4, 5, 6, 12, 14, 15) | 7 items (3, 7, 8, 9, 10, 11, 13) |
| 2.  | Geometry and Trigonometry | 3 items (20, 21, 23) | 5 items (16, 17, 18, 19, 22) |
| 3.  | Statistics and Probability | 5 items (24, 25, 26, 28, 29) | Item (27) |
| 4.  | Calculus | 1 item (30) | 5 items (31, 32, 33, 34, 35) |
|     | Amount   | 18 items | 17 items |
Based on Table 2, it can be seen that there are 18 items (51.43%) that have moderate difficulty and 17 items (48.57%) that are difficult. From 17 items that have difficult categories, each material tested has 1 test item that was the most difficult. This can be seen from the index of the level difficulty. The item for the most difficult algebra material is item 8.

The negation from the statement “If the dry season is prolonged, then all plants will dry out” is ....
A. If the dry season is not prolonged, then all plants will dry out.
B. If all plants will not dry out, then the dry season is prolonged.
C. The dry season is prolonged, and the plant will dry out.
D. The dry season is prolonged, but there is a plant that will not dry out.
E. If all plants will dry out, then the dry season is prolonged.

**Figure 1.** The most difficult item in algebra

The percentage of students who answered correctly in question number 8 was 7.3%. Item 8 is about math logic. To be able to answer item 8 precisely, students should understand the concept of the negation statement in the form of implication. The difficulty that may be faced by student is that student doesn’t understand this concept. This can be seen from the number of students who chose C choice as much as 61.2%. The possibility that students are only memorizing the negation of implication statements is a conjunction statement. Because it is the only memorization, so there is a misconception.

The most difficult item for geometry and trigonometry is number 18.

The image from point P (-3,2), which is reflected in the Y-axis, then continues to be translated T (4, -3) is....
A. (-7,5)  B.(-1,-1)  C. (1,-1)  D. (7,-5)  E. (7,-1)

**Figure 2.** The most difficult item in geometry

The percentage of students who answered item number 18 correctly is 11.9%. The question is an item about the composition of transformation. The transformation composition is a sequence of transformations that the operations are arranged by the composition rules. To solve this item, students must understand the composition rules in transformation. The sequence of the transformation in this item is \((x, y) \xrightarrow{Reflection \ of \ the \ y-axis} P'(-x, y) \xrightarrow{Translated \ by \ T(a,b)} P''(-x + a, y + b)\). The difficulty that students may be faced in this step is wrong in the transformation sequence. In addition, the student must understand the concept of reflection and translation. The result of the transformation composition in this item is \(P(-3,2) \xrightarrow{Reflection \ of \ the \ y-axis} P'(3,2) \xrightarrow{Translated \ by \ T(7,0)} P''(7,-1)\).

In statistic and probability material, there is only 1 difficult item that is item 27.

The number of ways to choose OSIS daily administrator, which includes the chairman, secretary, and treasurer of 8 candidates, is ....
A. 56  B. 108  C. 224  D. 336  E. 1344

**Figure 3.** The most difficult item in statistic and probability

The percentage of students who answered correctly in this item is 19.6%. The question is an item about permutation. The difficulties that may be faced by students in this item can be caused by several things. First, the student must understand the problem to determine the permutation or combination strategy that must be used. Second, the student doesn’t understand how to calculate with the permutation strategy.
The most difficult item in calculus is item 35.

Figure 4. The most difficult item in statistic and probability

The percentage of students who answered correctly in this item is only 1%. The question is about the application of definite integral. To solve this question, there are several steps that must be done. First, the student should understand the concept of the area that is limited by two curves. Therefore students must be able to draw to determine the area. Second, students must be able to determine the boundaries of the area formed by two curves. And the last step, the student should do the calculation of definite integral.

From the stages of solving this item, it can be predicted the cause of the difficulties faced by students is students’ confusing and difficult to combine several concepts.

Another parameter that determines the quality of the test is item discrimination. The discrimination index in this study is determined using the biserial point correlation index. The discrimination index of good items is greater or equal to 0.3 [6]. The discrimination item of mathematics National Standardized School Examination test is presented in table 3.

Table 3. Discrimination item of the Mathematics National Standardized School Examination

| Criteria | Category | Number of items |
|----------|----------|-----------------|
| ≥ 0.3    | Good     | 6               |
| < 0.3    | Not good | 29              |
| Amount   |          | 35              |

Based on Table 3, it can be seen that there are 29 not good items and only 6 good items. Therefore, this result can be said that the overall items of this instrument are not good. It means that the test cannot differentiate the low and high skilled participants.

The last parameter used to determine the quality of items is seen from the effectiveness of the distractor. According to Haladyna and Downing [13] stated that 5% of examinees must choose each distractor. This means that distractors that are less than 5% are ineffective distractors. Table 4 presents the classification of items based on the effectiveness of distractors.

Table 4. The classification of items based on the effectiveness of distractors

| Criteria | Category           | Number of items |
|----------|--------------------|-----------------|
| ≥ 5%     | Functioning distractor | 29              |
| < 5%     | Non-functioning distractor | 6               |
| Amount   |                     | 35              |

Based on Table 4, it can be seen that 29 items (82.86%) have good distractors because more than 5% of the examinees were selected, and 6 items (17.14%) have not good distractors. Therefore, it can be concluded that mathematics National Standardized School Examination of vocational school has a good level of effectiveness distractor because most items have effective distractors.

4. Discussion

The result of the qualitative analysis by proving the validity of the instrument obtained the Aiken (V) index of 0.94. It means that the mathematics National Standardized School Examination of vocational school is categorized as very valid. These results can be said that all test items are relevant to the standards of graduation competency, basic competencies, indicators, and cognitive levels. Therefore, these test kits have completed the validity requirements because Allen and Yen [12] stated that the validity of the instrument is the extent to which the measuring instrument is able to measure what should
be measured. The validity of an instrument is important to fulfill because validity is the most basic fundamentals in developing and evaluating a test [14].

Furthermore, an important thing that needs to be considered in the development of instruments is reliability. The reliability estimation of mathematics National Standardized School Examination of vocational school obtained an alpha coefficient of 0.362, which is categorized as moderate or low reliability. This is as stated by Retnawati [6] if the coefficient of a test score is low, then the reliability of the test is low. The reliability of the test is the degree of consistency of the test result. Thus, these kits of the test have a low level of consistency so that if this kit is used to measure the same thing at different times, then it is possible that the results will be different. In addition, reliability is related to measurement error. According to [6] if the reliability score gets smaller, then the measurement result will be even greater the error. There are several factors that can cause measurement errors including the characteristics of the instruments used themselves, for example, the preparation and implementation of measurements that do not follow the standard rules, the quality of the items in the instrument is not good, there was cooperation during the test, there were questionable instrument items, and the situation of examinees during the test such as participants who are tired both physically and psychologically [6].

The result of the analysis of the characters in terms of difficulty level in the mathematics National-Based School Examination, there were 17 difficult items and 18 moderate items. These results show that the difficulty level formulated by the test compiler is not appropriate with the level difficulty based on empirical analysis. It can be seen in the blueprint of this instrument kits, which only categorizes 7 items in level 3 or difficulty level. According to Stanley and Wang [15], this happened because the classification of the level difficulty by the test compiler was only based on intuition. Even though the items considered as difficult or easy items by the test compiler may not necessarily be felt difficult or easy by the examinees [16].

Based on the analysis of the difficulties described above, the researchers found that there were several possible causes of difficulties faced by students, namely the lack of students’ understanding toward the mathematic concepts, the students’ confusing and difficulties in combining one concept to another, error in applying procedure, and the students’ difficulties in identifying the useful information from test items. The lack of students’ understanding of the material tested was also found in the study conducted by Retnawati, Kartowagiran, Arlinwibowo, and Sulistyaningsih [17]. This is also supported by Rittle-Johnson, Schneider, and Star [18], which stated that students' weaknesses in basic mathematical ability were conceptual understanding. Even though the understanding of mathematical concepts is the most basic ability that students should have [19].

To overcome students’ difficulties can be done by strengthening students’ understanding of mathematical concepts. Understanding student concepts can be successful by improving the quality of learning because learning behavior can affect student achievement [20]. Teachers are expected to provide meaningful learning. Meaningful learning can be done by utilizing the context that is often encountered by students. This is in accordance with the resulting research by Ali [21], which concept understanding can be achieved by providing activities and something real that is around students. In addition, according to Bradley, Notar, Herring, and Eady, [22], meaningful learning can be implemented by linking concepts that have been known by students.

Improving learning can also be done by designing a learning trajectory. A learning trajectory is a part of a lesson plan. Based on the research result of Retnawati [23], a learning trajectory might assist students in understanding the mathematic concepts because Gravemeijer, Browsers, and Stephan [24] stated that with learning trajectory, learning becomes social aspect, continuous iterative cycle and offering the teacher learning theories that might be adjusted to students’ ability and reasoning level.

Another cause of students’ difficulty is an error in applying procedure. The cause of this difficulty was also found in the study of McCowan and McCowan [25], as many as 72.6% of students had difficulties in performing procedures in problem-solving. Basically, procedural ability and conceptual ability support each other [26]. This is also in line with Rittle-Johnson, Loehr, and Fyfe [27], which stated that improving procedural abilities would support the improvement of students’ conceptual abilities.
To improve procedural skills can be done by providing routine exercise with varied problems. According to Okigbo and Okeke [28] suggested that mathematical skills can be achieved by repetition so that students can solve various types of math problems. The provision of the various exercise might also increase the students' repertory of test items, so when they should complete test items with a similar or different concept, the students would have not any difficulties. In addition, the worked example is also suggested by Booth, Lange, Koedinger, and Newton [29] to improve students' procedural abilities, where the teacher can analyze student work procedures.

Based on the discrimination item, the mathematics National Standardized School Examination test has a low discrimination index that means the overall test items are not good. This was confirmed by the result of the analysis, which showed 82.86\% of the items are not good. According to Thorndike and Thorndike [30], low discrimination index is caused by several things, namely the level of difficulty of the item is too low or too high and there is a distractor that does not make sense, even though the difficulty level of the item is categorized as the good item. These distractors will make the item test easier for examinees to determine the wrong answer so that the item will be easier to answer [31].

The other parameter that can determine the quality of test items is the effectiveness of distractors. The percentage of items that have functioning distractors reaches 82.86\%, which means the distractors are chosen by participants equally which exceeds 5\%. It can be said that the mathematics National Standardized School Examination items have good quality. This is in accordance with the theoretical review of Kusaeri and Suprananto [32] which stated that in the good item, the distractors would be chosen equally by the test participants who answered incorrectly. Otherwise, in not good items, the distractor will be chosen unequally. However, the item whose distractor does not function must be corrected because it can provide an opportunity to answer correctly by guessing. In addition, according to Amelia and Kriswantoro [31], a distractor that makes no sense will make it easier for test participants to decide that the distractor is wrong, so they might answer correctly by guessing. And otherwise, the distractor that is too close to the correct answer will cause the item to be too difficult. Therefore, based on the research result of Santosos [33], it was concluded that the distractor’s function and the level of difficulty of the test items were related to each other so that in preparing a set of tests both of them needed to be considered.

5. Conclusion and Suggestions
From the qualitative analysis, it can be concluded that the mathematics National Standardized School Examination test is considered to be valid, which means that it is in accordance with what should be measured. Based on the result of quantitative analysis, in terms of reliability, the test has low reliability. For the level of difficulty of the overall items, it cannot be considered to be good. The difficulties experienced by students were suspected due to the lack of students’ understanding of the mathematics concepts, the students’ confusing and difficulties in combining one concept to another, error in applying procedure, and the students’ difficulties in identifying the useful information from the test items. From the discrimination item, the test can be categorized as not good, whereas in terms of the effectiveness of the distractor, this test is considered to be effective.

From this result of the study, there are a number of suggestions that can be made both for teachers and other researchers. First, the result of this study can be used as a basis for the compilers of test kits to improve the quality of tests. Second, the compiler of the test should know and understand in-depth the characteristics of the quality test, so that the items arranged are not too difficult and not easy. Third, the effectiveness of the distractors also needs to be considered for the test compilers so that the test developed is not easily answered by guessing. Fourth, from the results of the analysis of student difficulties, what the teacher should do is strengthening the students’ understanding, providing routine exercise with varied problems, and using a worked example in learning. Fifth, for other researchers, further research can be carried out by using characteristic analysis of tests based on item response theory for both the same or different test kits.
6. References

[1] Kemdikbud 2016 Salinan Permendikbud Nomor 23 Tahun 2016 Tentang Standar Penilaian Pendidikan (Jakarta: Kemendikbud)
[2] Ebel R L and Frisbie D A 1991 Essentials of Educational Measurement, no. 9. (New Delhi: Prentice-Hall of India)
[3] Roznowski M and Reith J 1999 Educ. Psychol. Meas. 59 pp 248–269
[4] Kartowagiran B, Munadi S, Retnawati H, and Apino E 2018 SHS Web Conf. 42 pp 1–6
[5] Razak N Bin A, Khairani A Z bin, and Thien L M 2012 Procedia-Social Behav. Sci. 69 pp 2205–2214
[6] Retnawati H 2016 Analisis kuantitatif instrumen penelitian (Yogyakarta: Parama Publishing)
[7] Mardapi D 2012 Pengukuran penilaian dan evaluasi pendidikan (Yogyakarta: Nuha Medika)
[8] Sutrisno H 2016 J. Ris. Pendidik. Mat. 3 p 162
[9] Adedoyin O O and Mokobi T 2013 Int. J. Asian Soc. Sci. 3 pp 992–1011
[10] Kemdikbud 2018 Salinan Permendikbud No 4 Tahun 2018 tentang Penilaian Hasil Belajar oleh Satuan Pendidikan dan Penilaian Hasil Belajar oleh Pemerintah (Jakarta: Kemendikbud)
[11] Guilford J P and Fruchter B 1978 Fundamental statistics in psychology and education (New York, NY, US: McGraw-Hill)
[12] Allen M J and Yen W M 1979 Introduction to measurement theory (California: Brooks/Cole Publishing Company)
[13] Haladyna T M and Downing S M 1993 Educ. Psychol. Meas. 53 pp 999–1010
[14] Anggriyani D 2018 Analisis kualitas butir soal ujian akhir semester mata pelajaran matematika kelas X SMA Muhammadiyah Kabupaten Bantul (UNY: S2 Thesis)
[15] Stanley J C and Wang M D 1968 Differential weighting: A survey of methods and empirical studies (Baltimore, MD: Johns Hopkins)
[16] Baker F B 2001 The basic of item response theory (USA: ERIC Clearinghouse on Assessment and Evaluation)
[17] Retnawati H, Kartowagiran B, Arlinwibowo J, and Sulistyaningsih E 2017 Int. J. Instr. 10 pp 257–276
[18] Rittle-johnson B, Schneider M, and Star J R 2015 Educ Psychol Rev 27 pp 587–597
[19] Retnawati H, Arlinwibowo J, and Sulistyaningsih E 2017 Int. J. New Trends Educ. Their Implic. 8 pp 28–41
[20] Pimta S, Tayruakham S, and Nuangchalerm P 2009 J. Soc. Sci. 5 pp 381–385
[21] Ali T 2011 Bull. Educ. Res. 33 pp 47–69
[22] Bradley J R, Notar C E, Herring D F, and Eady C K 2008 Asian Soc. Sci. 4 pp 60–65
[23] Retnawati H 2017 Olympiads in Informatics 11 pp 123–142
[24] Gravenmijer K, Browsers J, and Stephan M 2015 J. Res. Math. Educ. Monogr. 12 pp 51–66
[25] Mccowan R J and Mccowan S C 1999 Item Analysis for Criterion-Referenced Tests (New York)
[26] Nebesniak A L 2013 Math. Teach. 106 pp 354–358
[27] Rittle-Johnson B, Fyfe E R, and Loehr A M 2016 Br. J. Educ. Psychol. 86 pp 576–591
[28] Oikigbo E C and Okeke S O C 2011 Educ. Res. Rev. 6 pp 292–298
[29] Booth J L, Lange K E, Koedinger K R, and Newton K J 2013 Learn. Instr. 25 pp 24–34
[30] Thomdike R M and Thorndike-Christ T 2010 Measurement and evaluation in psychology and education. (Upper Saddle River, NJ: Pearson Education)
[31] Amelia R N and Kriswantonro 2017 J. Kim. dan Pendidik. Kim. 2 pp 1–12
[32] Kusaeri and Suprananto 2002 Pengukuran penilaian pendidikan (Yogyakarta: Graha Ilmu)
[33] Santoso A 2018 J. Pendidik. Mat. dan Sains 6 pp 158–168