The characteristics of mathematical material deepening test for the national examination of junior high school

Y Prasetya and Hartono
Deparment of Mathematics Education, Graduate School Program, Universitas Negeri Yogyakarta, Indonesia

E-mail: yogaprasetya.2018@student.uny.ac.id

Abstract. The Material Deepening Test (MDT) is a test to prepare students to face the National Examination in Indonesia. The MDT test results are used as an evaluation and consideration of teachers towards students’ readiness to face the National Examination. The MDT test developed based on blue print of the National Examination. This study aims to describe the characteristics of the MDT for National Examination of junior high school in mathematics subject in Sleman Regency of Indonesia. This study is an explorative descriptive using a quantitative approach. Data collected by documentation technique. The characteristics of MDT analyzed using the classical test theory. The results show that the test is valid with Aiken’s index of 0.87 and included in the high category of content validity. Reliability index of test is 0.70 in good category. There are 19 items have discriminant index in good category, 8 items in quite good category, and 11 items in not good category. There are 3 items have difficulty index in easy category, 20 items in moderate category, and 15 items in difficult category. The most difficult material items for students are lines and angles, where the indicator is determining the angle formed from the real problem.

1. Introduction
Assessment of learning outcomes in education is needed to support the sustainability of education itself. Assessment of learning outcomes can be used to improve the quality of education. The Indonesian government uses an assessment of learning outcomes, namely the National Examination (UN). UN aims to measure the achievement of competencies in particular subjects with the Competency Graduates standard. The purpose of the UN itself is used as a basis for unit education programs, assessment of entry into tertiary education, and guidance and assistance to education units in their efforts to improve the quality of education, Permendikbud No. 4 of 2018 about assessment of learning outcomes by the education unit and assessment of learning outcomes by the government [1].

To face the National Examination, each school prepares their students to achieve the expected Graduates Competency Standards. One of the efforts made by the education unit in Sleman Regency to achieve these goals is conducting a Material Deepening Test (MDT) before the UN is implemented. MDT is a test conducted for students to stabilize the material before facing a national exam. MDT is also used as an evaluation material for student learning outcomes in certain materials so that teachers in each school can see in full the readiness of students in facing the National Examination. The MDT is made by a group of teachers, both at the regency and province level through the Forum of Subject Teachers (MGMP) and the Forum of Principal Work (MKKS). In-depth Tests The material made includes the subject matter that will be tested. This test was made based on the blueprint UN.
Evidence of the validity of the Material Deepening Test used as a preparation to face the UN has not been done. MDT is important to be used as a reference to see the achievement of student competencies before facing the National Examination, the validity of the test will affect the accuracy of what will be measured. According to Allen and Yen [2], a valid test will be able to measure what should be measured. According to Miller, Linn, and Gronlund [3], validity is the adequacy and appropriateness of the interpretation and use of the results of the assessment. Validity will show the support of empirical facts and theoretical reasons for the analysis of test scores and also the accuracy of measurements [4].

Proving the validity of a test items can use criteria validity, content validity, and construct validity [5]. Retnawati [4] states that the content validity can be proven through experts who assess the relevance of each instrument then the assessment results are used to calculate the expert agreement index with the Aiken’s index or Gregory index. The construct validity can be proven through factor analysis, both exploratory and confirmatory. The validity of the criteria can be proven by knowing the magnitude of the correlation between the respondents' scores obtained by the instrument to the scores considered as criteria. Therefore, it is crucial to see whether the examination of the national exam material deepening test is proven valid or not? So that the results obtained can measure the achievement of student competencies through the test device used.

In addition to the validity of a test items, it is also necessary to analyze the characteristics of the test items to determine the quality of the test empirically [6]. According to Crocker and Algina [7], an outline analysis of the characteristics of test items can use the classical test theory approach and item response theory. Each of the item analysis approaches has advantages and disadvantages in analyzing the item used. According to Kartowagiran [8], classical test theory is more suitable for small respondents (approximately 100 respondents), whereas for large respondents (more than 200 respondents) suitable to use item response theory.

Classical test theory can analyze test items empirically including reliability, level of difficulty, and discriminant of a test item. The test is said to be reliable if the observed score has a high correlation with the actual score [2]. Where the reliability of a measuring instrument in the form of a value called the reliability coefficient, can be done with statistical calculations [4]. Reliable measuring instruments will also provide stable and consistent measurement results [9]. This means that a measuring tool with a high-reliability coefficient, which is used to measure the same thing at different times the results are the same or close to the same.

The level of difficulty or the so-called difficulty index (p). Difficulty index of the test items is useful to see how good the quality of test. If p approaches 0, then the item is too difficult, whereas if p approaches 1, then the item is too easy so it needs to be discarded because these items cannot distinguish the abilities of one student with other students [4]. The discriminant of a test item is to provide an overview of the test results in accordance with the actual abilities of students. Discriminant index is said to be good if it is greater or equal to 0.3 [10]. The discriminant index needs to be considered if the value is negative [11]. Then the test items cannot be included in the analysis so that the results of the analysis of the discriminant index obtained illustrate the compatibility with the actual abilities of students [4].

Based on the explanation and problem above, this study aims to describe the characteristics of mathematical Material Deepening Test for the National Examination of junior high school based on classical test theory. This research is expected to provide an overview of the analysis of the characteristics of items in the Material Deepening Test of Mathematics subject and can be used for future mathematical MDT development.
2. The Research Methods
This research was a descriptive exploratory study with a quantitative approach that aims to describe the characteristics of mathematical material deepening test for the national examination using classical test theory. The data collection was carried out through documentation of 40 items of A category of mathematical deepening test for the national examination worked by 115 Mts Students answer in Sleman regency. The form of the test was multiple choices with four answer choices.

The research consisted of five steps. The first step was making a blueprint of mathematical material deepening tests based on national examination (UN) blueprint. The second step was content validation. Content validation was done by four expert judgment of mathematics education. In the third step, the result of the test items analyzed by using classical test theory with the QUEST program. The fourth step, described the reliability, difficulty index, and discrimination index of MDT. The last step made conclusions about the results of the study viewed by the quality of the test items. The results obtained will use for further test improvement.

![Figure 1. Research Procedures.](chart)

3. Result and Discussion
3.1 Results of Validation of MDT
The validation of the Material Deepening Test uses content validity, validated by 4 (four) experts in the field of mathematics. Where the item is assessed in the form of compatibility between the subject and the indicator. After an assessment by each expert, the recapitulation results will be calculated by an expert agreement index (rater agreement) with the Aiken’s index. Aiken's [12] proposed item validity index is formulated as follows:

\[
V = \frac{\sum s}{n(c - 1)}
\]

where \(V\) is the rater agreement index regarding item validity; \(s\) score assigned to each rater minus the lowest score in the category used \((s = r - r_o, \text{ with } r\) is the score of the rater choice category and \(r_o\) is the lowest score in the scoring category); \(n\) number of raters; and \(c\) the number of categories rater can choose from.

The results of the calculation of index \(V\), an item or device can be categorized based on the index. If the index is less or equal to 0.4 \((V \leq 0.4)\) the validity is low, \(0.4 < V \leq 0.8\) is said to be of medium validity, and if greater than 0.8 \((V \geq 0.8)\) is said to be high or very valid [4]. The results of the recapitulation calculations from an expert agreement using the Aiken’s index are presented in Table 1.
Table 1. Expert agreement index.

| Items | Topics                      | V  | Category | Items | Topics                      | V  | Category |
|-------|-----------------------------|----|----------|-------|-----------------------------|----|----------|
| 1     | proportion                  | 0.69| medium   | 21    | spldv                       | 0.87| high     |
| 2     | rational numbers            | 0.81| high     | 22    | spldv                       | 0.94| high     |
| 3     | proportion                  | 0.69| medium   | 23    | pythagoras                  | 0.94| high     |
| 4     | integers                    | 0.56| medium   | 24    | square area                 | 0.94| high     |
| 5     | operation                   | 1   | high     | 25    | square                      | 0.75| medium   |
| 6     | powers of numbers           | 0.87| high     | 26    | similarity                  | 0.94| high     |
| 7     | social arithmetics          | 1   | high     | 27    | similarity                  | 0.94| high     |
| 8     | social arithmetics          | 0.87| high     | 28    | similarity                  | 0.87| high     |
| 9     | sequence and series         | 1   | high     | 29    | triangles                   | 1   | high     |
| 10    | sequence and series         | 0.94| high     | 30    | lines and angles            | 0.94| high     |
| 11    | algebraic forms             | 0.94| high     | 31    | circle                      | 0.75| medium   |
| 12    | algebraic forms             | 0.87| high     | 32    | plane geometry              | 1   | high     |
| 13    | algebraic forms             | 0.75| medium   | 33    | plane geometry              | 0.62| medium   |
| 14    | proportion                  | 0.94| high     | 34    | solid geometry              | 0.81| high     |
| 15    | sets                        | 0.94| high     | 35    | solid geometry              | 1   | high     |
| 16    | sets                        | 1   | high     | 36    | statistic                   | 0.75| medium   |
| 17    | sets                        | 0.69| medium   | 37    | statistic                   | 0.75| medium   |
| 18    | sets                        | 0.81| high     | 38    | statistic                   | 0.94| high     |
| 19    | linear equation             | 0.81| high     | 39    | probability                 | 0.87| high     |
| 20    | linear equation             | 1   | high     | 40    | probability                 | 1   | high     |
|       | **Average (V)**             |     |          |       |                             | 0.87| high     |

In Table 1 above, MDT Aiken’s index is in the medium and high category. 10 items of medium category and 30 items of high category. The Aiken’s index of the average device or test item obtained, which is 0.87, can be said to have a high category of content validity.

3.2 Reliability
According to Miller, Linn and Gronlund [3] good and poor reliability of the instrument can be determined using the categories contained in Table 2.
To see the reliability of the instrument we can see *internal consistency* in the results of the QUEST program, shown in the Figure 2.

![Figure 2. Results of Reliability Using the QUEST Program.](image)

It can be seen that the internal consistency is 0.70 which means that the reliability of the MDT test. Reliability index is good category. This means that the reliable of the test can be categorized as good in measuring students' actual abilities.

### 3.3 Characteristics of MDT Based on Classical Theory Tests

The characteristics of MDT items of National Examination mathematics material were analyzed using the QUEST program. The analysis results for the test items are presented as follows.

#### 3.3.1 Difficulty Index

Difficulty Index ($p_i$) in the classic item analysis is the average scale of test takers who correctly answers an item. In the scale analysis used is the proportion of correct answers. The magnitude of the level of difficulty ranges from 0 to 1. The level of difficulty of an item in this study can be categorized into three, namely easy ($p_i > 0.70$), moderate ($0.30 \leq p_i \leq 0.70$) and difficult ($p_i \leq 0.30$) [3].

Based on the results of the QUEST program and with the difficulty category items above, the results of difficulty index is presented in Table 3.

| Category   | Items          |
|------------|---------------|
| Easy       | 1, 14, 38     |
| Moderate   | 2, 3, 4, 5, 6, 7, 8, 9, 11, 13, 16, 18, 21, 22, 23, 25, 26, 29, 33, 40 |
| Difficult  | 10, 12, 15, 17, 19, 20, 27, 28, 30, 32, 34, 35, 36, 37, 39 |

In the table above, it can be found that there are only 3 items which are easy or equal to 7.8%, and 20 items which are moderate or equal to 52.6% and there are 15 items that are difficult or around 39.4%. If seen from the QUEST results on the Estimates Tresholds item, the most difficult and many wrongs in number 30 where 19 students answered correctly and 96 students answered incorrectly.

#### 3.3.2 Discrimination Index

Discrimination Index ($r_{pbis}$) is said to be good and acceptable if it is greater or equal to 0.3, is quite good and needs to be fixed $0.2 < r_{pbis} < 0.3$ and is not good and is not acceptable $r_{pbis} < 0.20$ [10]. More detailed
information related to the characteristics of test items using classical theory using the QUEST program is outlined in the Table 4.

**Table 4. The Discrimination Index of Items.**

| Items | $r_{pbi}$ | Items | $r_{pbi}$ |
|-------|-----------|-------|-----------|
| 1     | 0.29      | 21    | 0.39      |
| 2     | 0.26      | 22    | 0.06      |
| 3     | 0.3       | 23    | 0.32      |
| 4     | 0.35      | 24    | -0.02     |
| 5     | 0.18      | 25    | 0.37      |
| 6     | 0.31      | 26    | 0.24      |
| 7     | 0.52      | 27    | 0.1       |
| 8     | 0.25      | 28    | 0.53      |
| 9     | 0.47      | 29    | 0.27      |
| 10    | 0.19      | 30    | 0.03      |
| 11    | 0.28      | 31    | -0.04     |
| 12    | 0.17      | 32    | 0.29      |
| 13    | 0.2       | 33    | 0.38      |
| 14    | 0.29      | 34    | 0.252     |
| 15    | 0.24      | 35    | 0.48      |
| 16    | 0.38      | 36    | 0.11      |
| 17    | 0.48      | 37    | 0.11      |
| 18    | 0.1       | 38    | 0.1       |
| 19    | 0.3       | 39    | 0.36      |
| 20    | 0.33      | 40    | 0.36      |

Based on Table 4, there are 12 items with not good discrimination Index, especially numbers 24 and 31 which have negative discrimination values (cannot distinguish between students with high ability and students with low ability) so that they are eliminated from being included in the analysis [4]. The results of an analysis of 38 questions can be seen that a good discrimination index is 19 test items, then quite good discrimination index is 8 test items, and not good discrimination index is 11 test items.

### 3.4 Discussion

The results of the analysis to prove the validity of the contents of the questions tested on MDT which have been validated by experts show that relevant with the average score of the Aiken's index obtained 0.87 included in the high category of content validity. This means that the suitability of the test items with the indicators to be tested is relevant.

The results of the analysis of the characteristics of test items using the classical test theory with the help of the QUEST program show the results of the level of difficulty of items as many as 3 easy categorized items, 20 moderate categorized items and 15 difficult categorized items. If it is presented, it can be obtained that there are only 3 easy items, namely 7.8%, and 20 items which are moderate or equal to 52.6% and there are 15 difficult items or around 39.4%. Items that are difficult for students to work on almost 40% of the total, this clearly shows the level of difficulty of mapping out the level of difficulty of the questions that are not good if given to students with different abilities. In Karnoto's opinion [13], the difficulty index of difficult questions should be in the range of 15% so that the items tested can represent each ability possessed by students. This means that the difficulty level of the difficult questions given does not exceed 15% of the total. The question numbers included in the difficult category are numbers 10, 12, 15, 17, 19, 20, 27, 28, 30, 32, 34, 35, 36, 37, 39. When seen in the QUEST results on the Estimates Thresholds item, the most difficult and many wrong in number 30 with a percentage of 16.52% of students who managed to answer correctly. This means that there are as many
as 19 students who can answer correctly from 115 students who took the MDT test. This item number 30 is classified as very difficult. The question number 30 is

The angle formed by the long needle and the short needle when the clock shows 03.25 is …

a. $52.5^\circ$
b. $47.5^\circ$
c. $30.0^\circ$
d. $12.5^\circ$

The subject matter in the matter above is about lines and angles, where the indicator is determining the angle formed from the real problem. There are several factors that make the matter of lines and angles categorized as difficult. First, the terminology used or the size of the numbers involved in the matter of lines and angles [14]. The measurement angle must be more than 90 degrees to help students or use angles that can be realized by students [15]. Second, understanding students' concepts of material lines and angles. According to the study of I.L. Irsal et al [16], almost all students have a low instrumental understanding of the topic lines and angles on different real problems and have low relational understanding. Third, students' problem-solving abilities. Problem-solving ability becomes important for students because it will affect the development of students' conceptual understanding [17]. The most important factor is how often the teacher gives students real practice exercises related to lines and angles and how to convey material lines and angles to be meaningful [18]. So students feel familiar with the item lines and angles during the test.

Actually, there are many factors that influence why many of the questions given are difficult to categorize, both in terms of the question maker or students as tested in the test. For example, according to the makers of questions that are categorized as moderate or easy, but for participants who work in the difficult categories or can also be called based on intuition to classify the difficulty level of the questions made [19], so are the readiness of test participants in working on the test and mastery of the material learned.

The discrimination index of MDT which consisted of 40 analyzed items, there are two items of negative values, especially numbers 24 and 31, so that they will not be included in the analysis characteristic. The discrimination index that is negative value must be discarded or cannot be included in the analysis because the item cannot distinguish high-ability students from low-ability students [11], so there are 38 items analyzed for MDT characteristics items. There are 19 items out of 38 items that have good discrimination index, 8 items out of 38 items of quite good quality, and 11 items of 38 that are of not good quality. When presented, many items of good quality are only 50% of the total, and sufficient quality is 22%, and there are 28% of not good quality. Half of the overall items are categorized enough to be able to distinguish the ability of students who are smart and less clever. However, it is also important to underline that almost 30% of the test items that show not good discrimination index, and there are 5% of items that have not good discrimination index that it is not worth testing. According to Thorndike [20], some things that cause not good discrimination index are the difficulty level of the item that is too low or too high or there is an unreasonable deception, even though the level of difficulty of the item is in a good category.

Improvements are needed for some test items that are categorized as not good discrimination index and high difficulty so that the test can truly measure students' actual abilities. It should also be noted that distractor test items is good and makes sense so that students indeed answer from working on test questions with answers that may be true, not just coincidence. So discrimination index, the level of difficulty, and distractors are also interrelated to determine a good test quality.

4. Conclusion and suggestion
The results showed that the content validation (Aiken’s index) MDT test was at a score of 0.87 included in the high category of content validity. There are two item that have not good discrimination index with negative scores so they are not included in the analysis. The reliability of the test obtained was
There are 3 items have difficulty index in easy category, 20 items in moderate category, and 15 items in difficult category. The items were very difficult for students, especially in the material of lines and angles, where the indicator is determining the angle formed from the real problem. The discrimination index of the categorized test items is quite good with a percentage of 71% (consisting of 19 good items and 8 quite good categories), but there is still not good discrimination index category of 29% (consisting of 11 items).

Based on the results of classical theory analysis, it consists of the level of difficulty, discrimination index, and the reliability of MDT test items. Items that contain a not good discrimination index must be repaired, and too many difficulties items must be made evenly does not exceed 15% of the total items test to produce a good quality Material Deepening Test (MDT).

The implication of this research when making a test item, and it is important to consider the discrimination index, the level of difficulty, and the distractor of the answers so that the test item can be good and measure the expected competency abilities. Second, further analysis of MDT characteristics using the item response test approach to obtain more accurate results and can be considered for improving the test in order to have a good test quality.

References
[1] Kemendikbud 2018 Permendikbud No 4 Tahun 2018 tentang Penilaian Hasil Belajar (BSNP: Jakarta)
[2] Allen M J and Yen W M 1979 Introduction Of Measurement theory (California USA: Brooks/Cole Publishing Company).
[3] Miller M D, Linn R L, Gronlund N D 2009 Measurement and Assessment in Teaching (New Jersey USA: Pearson).
[4] Retnawati H 2016 Analisis Kuantitatif Instrumen Penelitian: Panduan untuk peneliti, mahasiswa, dan psikometrian (Yogyakarta: Parama Publishing)
[5] Lawrence, M R 1994 Question to ask when evaluating test (Eric Digest)
[6] Siri A dan Freddano M 2011 Procedia - Soc. Behav. Sci. 29 188–97
[7] Croker L and Algina J 1986 Introduction to classical and modern test theory (New York: Holt, Rinehard and Winston Inc)
[8] Kartowagiran B 2012 Penulisan Butir Soal (Universitas Negeri Yogyakarta)
[9] Mehrens W A and Lehmann I J 1973 Measurement and evaluation in education and psychology (New York: Holt Rinehart and Wiston, Inc)
[10] Mardapi D 2012 Pengukuran penilaian & evaluasi pendidikan (Yogyakarta: Nuha Litera) pp 38–45
[11] Frisbie D A 2005 Measurement 101: Some fundamentals revisited. Educational Measurement (Issues and Practice) pp 21-28
[12] Aiken L R 1980 SAGE Journal 40
[13] Arikunto S 2002 Dasar-dasar Evaluasi Pendidikan (Edisi Revisi) (Jakarta: Bumi Aksara)
[14] Foxman D and Ruddock G 2014 Concepts and Skills: Line Symmetry and Angle
The Mathematical Association [online] http://www.jstor.org/stable/30216196
[15] Halim N D A, Han O B, Abdullah Z and Yusup J 2015 AIP Conf. Proc. 1660
[16] I L Irsal et al 2017 J. Phys.: Conf. Ser. 895 012073
[17] NCTM 2010 Why is Teaching with Problem Solving Important to Student Learning [Online] http://www.nctm.org/news/content.aspx? id=25713
[18] N S Utami et al 2017 J. Phys.: Conf. Ser. 895 012037
[19] Stanley J C, Wang M D 1968 Differential weighting: A survey of methods and empirical studies (Baltimore, MD: Johns Hopkins)
[20] Thorndike R M 2005 Measurement and evaluation in psychology and education