ANALYSIS OF MATHEMATICS MID TEST QUESTIONS FOR CLASS VII SMP EDUCATION 21 KULIM

Anisa Putri\textsuperscript{1*}, Yenita Roza\textsuperscript{1}, Maimunah\textsuperscript{1}
\textsuperscript{1}Riau University
Email: aputri02.ap@gmail.com

DOI: http://dx.doi.org/10.26418/jpmipa.v13i1.44910

Abstract
This study aims to analyze the items for the mathematics middle exam at SMP Education 21 Kulim grade VII for the 2020/2021 academic year in terms of the level of validity, reliability, level of difficulty, and discriminating power of questions. The research method used is descriptive quantitative. The data is in the form of 20 essay questions which are analyzed using Anates software Version 4.0. The results of the analysis of the level of validity are stated to be not good because overall it has a valid percentage of <70%. From the level of reliability, it is stated that it is not good because it has a reliability of <0.7. From the level, it is stated that the difficulty level is not good because <70% of the total number is classified as moderate. From the power of distinction, it is stated that the overall question has good distinguishing power because <70% of the total questions are classified as good. The solution is that these questions are not used and revisions are made. Thus, based on the test quality indicators, it can be said that the final semester tests are not of high quality.

Keywords: Quality Item analysis, validity, Anates Version 4.0

INTRODUCTION
The development of educators or teachers as one of the main educational figures has an important role in the progress of education. Through the educational process, a person will gain knowledge, understanding and ways of acting according to Muhibbin (Fariza et al., 2019). The learning process in schools is closely related to quality education. According to Aunurrahman (Suprapt, 2020) the success of learning process is the peak of all activities carried out by teachers and students which means that whatever teacher activities, starting from designing learning, selecting and determining materials, approaches, strategies and learning methods, choosing and determining evaluation techniques, all of which are directed to achieve student learning success. Based on this, teachers are required to be able to carry out and process...
education according to their education background. The teacher's job is not only about teaching but also providing a policy for assessing learning outcomes (Sa’idah, 2017).

In the learning process, the teacher acts as an intermediary in providing knowledge and changing behavior in children. One of the teacher's duties in education is to evaluate learning outcomes.

Evaluation is one of activities carried out by teacher to determine the extent to which students understand the learning material according to Sapta (Kurniasi et al., 2020). Evaluation is also a process of collecting basic data, reviewing and making decisions from the results of data studied. Activities in this evaluation have been regulated by the Law of Republic of Indonesia No. 20 of 2003 concerning the National Education System Chapter XVI Article 58 paragraph 1 which states: evaluation of student results is carried out to monitor the process, progress, and continuous improvement of student learning outcomes (Depdiknas, 2003). Learning outcomes evaluation activities are carried out by educators to monitor the process, progress, achievement and continuous improvement of student learning outcomes according to Sukardi (Erawati, 2018).

Learning without evaluation activities will lose meaning. This is because teachers cannot obtain important information about the level of objectives achievement, mastery of learning materials, strengths, weaknesses of students in learning, as well as the weaknesses and strengths of teachers in the learning process developed. Evaluation is considered important and is a teacher's routine job, but in reality this evaluation system in learning is meaningless without problems. In the form of learning evaluation test is carried out. According to Anas the function of test in general is as measuring tool to measure the success of teaching programs and as a measuring tool for the development of student progress (Achadah, 2019).

Test is a measuring tool to obtain information on student learning outcomes that require correct or incorrect answers. The results of test or evaluation are common measuring tool used to determine students' understanding of material that has been delivered. In addition, from the questions used, it can be seen whether the questions can measure the curriculum objectives that have been set or not, so that the results can be used as benchmark for the implementation of learning objectives (Hamimi et al., 2020).

Based on this, a teacher must be accurate to determine the measuring instrument that will be used to see the level of success of learning process carried out and to measure the level of student understanding. This is in accordance with what was stated by Bukhari (2005), namely: "The tools used by teachers to collect data about students, were chosen really beforehand to get information about a child's intelligence". The first step that must be taken by the teacher before taking the test is to test the quality of the questions.

Item analysis aims to determine the quality of each item that will be tested on students. By analyzing these items, it can be seen the quality of good and bad items as well as items that
were revised, entered into the question bank, or discarded (Oktanin, 2015).

Analysis of item quality is a step that must be taken to determine the quality degree of test, both the test as a whole and the items that are part of the test. If the test used by teacher is not good, then the results obtained are certainly not good according to Arifin (Iskandar et al., 2017). A good test question is a question that can provide precise information, so that it can determine students who have or have not mastered the material taught by the teacher (Syafitri, 2020). This is in line with Rahmasari's opinion that test questions are said to have good quality if they are in accordance with the curriculum, meet the requirements for aspects of validity, reliability, and high discriminatory power, moderate level of difficulty and can measure the achievement of student competencies (Rahmasari, 2016).

Information was obtained from the results of interviews conducted with mathematics teachers at SMP Education 21 Kulim. During the implementation of both mid-test and end-semester examinations, a thorough item analysis was not carried out. It is because the time is not enough and lack of understanding to do this. This is also evidenced by the results of study by (Muhson et al., 2013) which shows that the willingness and ability of teachers to conduct item analysis is still low, both manually and using the help of item analysis programs.

Therefore, the quality of questions made still needs to be questioned, whether it meets the criteria as a good question or not. In order to find out a good test, it is necessary to do an item analysis, so that the items that are already good and the questions that are still not good will be known. If the item is good, then it can be said that the test is also good (Istika et al., 2019).

With this, the teacher can perform item analysis using the application. One of the free and easily available software is ANATES (Test Analysis), because it is very practical, fast and easy to understand because the application process uses Indonesian. The ANATES V4 application is very easy to use and can help in analyzing questions (Sannova et al., 2017).

Along with the rapid development of technology in all areas of life, including in the field of education, a program is created that is able to facilitate teachers and researchers in analyzing the quality of question, one of which is the Anates V4 program. This program was developed and designed to analyze items, both in the form of descriptions and in multiple-choice forms that have the ability to calculate scores (original and weighted), calculate reliability, classify subjects into upper and lower groups, calculate discriminatory power, calculate difficulty level, calculate correlation item score with the total score, and determine the quality of distractors (Sari et al., 2014).

Item analysis currently has an important role in the development of instruments, so many researchers have conducted research on item analysis, some of which analyzed items, the research conducted by Rina Irawati (2020) that it already has a good category with sufficient discrimination, all alternative answers are effective, and the level of difficulty is moderate (Irawati et al., 2020). Another research by Wulan Ni'Matus Rohmah in 2017 with the title
"Analysis of Evaluation Items for the End of Odd Semester Final Exams for Class IV Social Sciences in the Mojoroto District, Kediri City for the 2016/2017 Academic Year" using the ITEMAN 3.0 application and the results shown are classified as good questions (Wulan, 2017). Relevant research was also carried out by Putri (2013) with the title "Analysis of Final Examination Instruments for Science Subject Semester at State Junior High Schools in Ngawi Regency" the results of research conducted using Microsoft Excel, namely the quality of questions with not good category (Irawati et al., 2020)

Based on the above background, the researchers were motivated using the Anates V4 program to conduct research on the Quality Analysis of Mathematics Questions for the Mid-Semester Odd Examination for Class VII Year 2020/2021 in the form of essay questions in terms of level of validity, reliability, discriminating power, and level of difficulty of questions.

METHODS

This research technique uses a quantitative descriptive type of research to determine the quality of odd semester math questions for class VII SMP Education 21 Kulim in 2020/2021. 20 essay questions was analyzed. Research instrument is used to collect research data and used documentation. Analysis by documentation using question and answer sheet.

The criteria for drawing conclusions for the quality of a good item based on the level: validity, reliability, item difficulty, and distinguishing power according to Mardapi (Wulan, 2017) :

a. Validiti
1) A question is said to have good validity if it has a percentage of 70% of the questions are valid from the whole question.
2) Questions are said to have poor validity if they have a percentage of <70% valid questions from all questions.

b. Reliability
1) A question is said to have high reliability if it has a reliability ≥ 0.7
2) The question is said to have poor reliability if it has a reliability < 0.7

c. Difficulty level
1) A question is said to have a good level of difficulty if 70% of the total number of questions is classified as moderate
2) The question is said to have a poor level of difficulty if < 70% of the total number is classified as moderate

d. Power Difference
1) A question is said to have good discriminating power if 70% of the total number of questions is classified as good
2) Questions are said to have poor discriminatory power if < 70% of the total

RESULTS AND DISCUSSION

The results of the quality analysis of the questions in the form of 20 essay questions were carried out using the Anates Version 4.0 program which were divided into the level of validity, level of reliability, difficulty level of questions and level of distinguishing power as follows:
**Item Validity**

According to Sugiyono, validity is a degree that shows what is being measured from a test and can be said to be valid if data or information is in accordance with the actual situation (Hery et al., 2015). A test is valid according to Sukardi if the coefficient of 0.5 is accepted, on the other hand if there is another similar prediction test and has a higher coefficient then the coefficient of 0.5 is not accepted, then a valid test has a minimum coefficient of 0.5 (Elviana, 2020). According to Azwar, it can be said that an item is valid if a validity coefficient can be considered satisfactory with the coefficients obtained ranging from 0.30 to 0.50 (Azwar, 2008). In this study, the results of the analysis using the Anates version 4.0 program can be described in the following table:

| No | Correlation | Significance | Inf |
|----|-------------|--------------|-----|
| 1  | 0.025       | –            | TV  |
| 2  | 0.421       | –            | TV  |
| 3  | 0.411       | –            | TV  |
| 4  | 0.469       | Significant  | V   |
| 5  | 0.462       | Significant  | V   |
| 6  | 0.546       | Significant  | V   |
| 7  | 0.568       | Very Significant | V   |
| 8  | 0.487       | Significant  | V   |
| 9  | 0.282       | –            | TV  |
| 10 | 0.366       | –            | TV  |
| 11 | 0.510       | Significant  | V   |
| 12 | 0.216       | –            | TV  |
| 13 | 0.401       | –            | TV  |
| 14 | 0.312       | –            | TV  |
| 15 | 0.772       | Very Significant | V   |
| 16 | 0.633       | Very Significant | V   |
| 17 | 0.700       | Very Significant | V   |
| 18 | 0.305       | –            | TV  |
| 19 | 0.135       | –            | TV  |
| 20 | 0.388       | –            | TV  |

Based on the table above with the information V = valid and TV = invalid, it can be seen that there are 45% of the questions declared valid in questions number 4, 5, 6, 7, 8, 11, 15, 16, and 17. There are 55% of questions declared invalid on questions number 1, 2, 3, 9, 10, 12, 13, 14, 18, 19, and 20. Therefore, seen as a whole, according to Mardapi, the question was declared to be not good because it had a percentage of <70 questions that valid from all (Wulan, 2017).

From table 1 it can be seen that question number 1 is a question that is not significant or has the lowest correlation and question number 15 is a question that is very significant or has the highest correlation value and. the questions are:
Thus, according to Astuti's research, follow-up can be done on the results of item analysis, namely a) Invalid items are declared as invalid questions and should be discarded, but if they are to be reused, they should be revised. b) Valid question items can be reused and entered in the question bank (Astuti, 2020).

**Item Reliability**

Reliability according to Sugiyono is an index that shows the extent to which a measuring instrument can be trusted (Hery et al., 2015). According to Zainal, one of the characteristics of a question that has high reliability is if the item consists of many items with valid categories. In addition, the high and low reliability index is influenced by several factors, namely the length of the test, the distribution of scores, the level of difficulty, and objectivity (Astuti, 2020). In this study, the results of the analysis of questions using the Anates program version 4.0 can be seen in the following picture:

![Figure 1. Results of reliability analysis of questions](image)

Based on Figure 1 above, the reliability result of question is 0.61 where according to Mardapi that the question is declared to be not good because it has a reliability <0.7 (Wulan, 2017). According to Agustiana's research, the low reliability coefficient of questions is caused by the difficulty of the test. One of the factors that affect reliability is test difficulty, tests that are too easy or too difficult for students tend to produce low reliability (Agustiana et al., 2018).

**Item Difficulty Level**

Level analysis is examining the test questions in terms of difficulty so that it is known which questions are included in the very easy, easy, medium and difficult categories. Questions are declared good, namely questions that are not too easy and not too difficult. Difficulty level of questions is seen from the ability of students to answer them, not from the point of view of teacher as question maker. The results of the analysis of questions using the Anates version 4.0 program on the difficulty of the items are shown in the following table:

| No | Diff. Level (%) | Inf. |
|----|----------------|------|
| 1  | 96.00          | Easy |
| 2  | 80.00          | Easy |
| 3  | 90.00          | Easy |
| 4  | 84.00          | Easy |
| 5  | 76.00          | Easy |
The table above shows question difficulty level that there are 90% of items stated in easy criteria on questions number 1, 2, 4, 3, 5, 6, 7, 9, 10, 11, 12, 13, 14, 16, 17, 18, 19 and 20 and there are 10% of items stated in the moderate criteria in questions number 8 and 10. From the results of the overall analysis, according to Mardapi, the questions are stated to have a poor level of difficulty because <70% of the total number is classified as moderate (Wulan, 2017). This is in accordance with Nugraha's research that the items included in the difficult and easy categories indicate that in terms of material, the items are not representative of the material that has been taught.

From table 3, it can be seen that question number 8 is a question in the medium category and question number 13 is a question in the easy category. The question is:

| No. | Question | Difficulty | Inf. |
|-----|----------|------------|------|
| 8   | Determine the result of 0.893 + 12.642! | Easy       | 72.00|

Thus, what can be done on the results of item analysis are as follows (Nugraha, 2017):

a) Items based on the results of analysis are included in the good category (moderate difficulty), can be directly recorded in the question bank.

b) Question items in 'too difficult' category have three possible follow-ups that can be done, namely: First, the question items are discarded. Second, it is re-checked, researched and traced the causes of items that are difficult to work on. After that, repairs are made so that the items can be reused. The third can be used in tests that have a strict nature (selection tests) so that they can be stored in a separate question bank.

c) Items in the easy category have three possible follow-ups, namely: first, the item is discarded. The second is
rechecked, researched and traced to find out the factors causing the questions to be answered correctly by almost all students. Once known then to be corrected. The third can be used in loose tests.

**Power Difference**

Distinguishing power is the ability of question to distinguish the ability of students, both high-ability students and low-ability students (Putri, 2019). This is in accordance with Sundayana's opinion that discriminatory power is the ability of items to be able to distinguish students with high abilities from students with low abilities (Sundayana, 2016). In this study, the results of questions analysis using the Anates version 4.0 program are shown in the following table:

| No | t   | DP (%) | Inf.   |
|----|-----|--------|--------|
| 1  | 0.00| 0.00   | Bad    |
| 2  | 1.63| 40.00  | Sufficient |
| 3  | 1.00| 20.00  | Bad    |
| 4  | 1.73| 32.00  | Sufficient |
| 5  | 2.33| 48.00  | Good   |
| 6  | 1.91| 40.00  | Sufficient |
| 7  | 1.73| 32.00  | Sufficient |
| 8  | 1.49| 28.00  | Sufficient |
| 9  | 1.00| 20.00  | Bad    |
| 10 | 1.63| 40.00  | Sufficient |
| 11 | 1.63| 40.00  | Sufficient |
| 12 | 1.63| 40.00  | Sufficient |
| 13 | 1.00| 20.00  | Bad    |
| 14 | 1.00| 20.00  | Bad    |
| 15 | 8.57| 84.00  | Very Good |
| 16 | 3.58| 32.00  | Good   |
| 17 | 3.21| 48.00  | Good   |
| 18 | 0.43| 28.00  | Sufficient |
| 19 | 1.00| 20.00  | Bad    |
| 20 | 1.63| 40.00  | Sufficient |

Based on table 5 above, it shows that there are 30% of the items stated in the category of poor discriminating power on questions number 1, 3, 9, 13, 14, and 19, there are 50% of the items stated in the category of sufficient discriminating power on questions number 2, 4, 6, 7, 8, 10, 11, 12, 18, and 20, there are 15% of the items stated in the category of discriminatory power both in item numbers 5, 16 and 17, and 5% of the items stated in the category of good discriminating power. Once on question number 15. From the results of the analysis, it can be seen that as a whole the question has a poor distinguishing power because <70% of the total number is classified as good (Wulan, 2017).

This is in line with Arifin's opinion (Wati, 2015) that the higher the discriminating power coefficient of
an item, the more capable the item will be to distinguish between high-ability students and low-ability students. The following are examples of questions that are declared bad, sufficient, good, and very good.

| Category     | No. Que | Question |
|--------------|---------|----------|
| Bad          | 01      | The temperature inside a refrigerator changes from -20° to -27°. Write down the changes that occur in the temperature of the refrigerator. |
| Sufficient   | 06      | If $8a = 40$ and $-15 + b = -3$, determine $a + b$ ! |
| Good         | 05      | Determine the result of $(3^3 + 3^2)^4$ ! |
| Very Good    | 15      | Prove that 1.2666 is a rational number! |

Thus, what can be done on the results of the analysis of the differentiating power of the items, namely correcting or removing items in the bad category, correcting the items in the sufficient category and entering the items into the question bank in the good category. This is in accordance with Anetha's research, namely that the items in the sufficient, good and very good categories can be included in the question bank so that they can be reused on the condition that improvements are made to items in the sufficient category. Questions that have poor discriminating power can be discarded or repaired so that they can be reused (Anetha, 2019).

CONCLUSION

The results of the quality analysis of the mid-semester questions of SMP Education 21 Kulim in the city of Pekanbaru in class VII using the Anates Version 4.0 program can be concluded as follows. Judging from the validity that there are 45% of the items declared valid and 55% of the items declared invalid. Thus, the overall question is declared to have a poor percentage because <70% of the total valid amount.

In terms of reliability, the reliability result of question is 0.61 where the item is declared to be not good because it has a reliability <0.7. Judging from the level of difficulty, there are 90% of the include in the easy criteria and 10% of the items into the moderate criteria. Thus, the question can be stated to have poor level of difficulty because <70% of the total number is classified as moderate. Judging from the differentiating power, there are 30% items in the category of poor distinguishing power, 10% of the items in the category of sufficient distinguishing power, 15% of items in the category of good distinguishing power and 5% items in the category of very good distinguishing power. Thus, overall the question has a poor discriminatory power because <70% of the total number is classified as good.

Thus, based on indicators of validity, reliability, level of difficulty, and discriminating power, it can be concluded that the mid-test questions for mathematics subjects for SMP
Education 21 Kulim class VII students have poor quality.

REFERENCES
Achadah A. (2019). Evaluasi Dalam Pendidikan Sebagai Alat Ukur Hasil Belajar. *Jurnal An-Nuha* 6 (1).

Agustiana M, Mayrita H, & Muchti A. (2018). Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Bahasa Indonesia Kelas VII. *Jurnal Ilmiah Bina Edukasi*. 11(1), 26-35.

Arikunto, Suharsimi. (2010). *Dasar-dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara.

Anetha L. F. & nTilaar H. (2019). Analisis Butir Soal Semester Ganjil Mata Pelajaran Matematika pada Sekolah Menengah Pertama. *Jurnal Pengukuran Psikologi dan Pendidikan Indonesia* 8 (1), 57-68.

Astuti. (2020). Analisis Soal Ujian Akhir Semester Genap Mata Pelajaran Matematika Siswa SD Negeri 05 Binuang Ajaran 2017/2018. *Jurnal Pendidikan Tambusai* 04(1), 867-880.

Azwar, Saifuddin. (2008) *Reliabilitas dan Validitas*. Yogyakarta: Pustaka Pelajar.

Bukhari. (2005). *Teknik-teknik Dasar Evaluasi Pendidikan*. Bandung: Jemmars.

Depdiknas. (2003). *Undang-undang RI No.20 tahun 2003. Tentang sistem pendidikan nasional.

Elviana. (2020). Analisis Butir Soal Evaluasi Pembelajaran Pendidikan Agama Islam Menggunakan Program Anates. *Jurnal Mudarrisuna* Vol.10 No.2 hal 58-74.

Erawati, K. N. (2018). Analisis Tes Penilaian Pencapaian Kompetensi Pada Mahasiswa. *Jurnal Penjekora* 5 (2).

Fariza D, & Yeyenti N. (2019). Pengaruh Interaksi Teman Sejawat Terhadap Disiplin Disekolah Peserta Didik Kelas VII dan VIII SMP Muhammadiyyah Palangka Raya. *Jurnal Suluh Bimbingan dan Konseling* 4(2).

Hamimi, L., Zamharirah, R., & Rusydy, (2020). Analisis Butir Soal Ujian Matematika Kelas VII Semester Ganjil Tahun Pelajaran 2017/2018. *Mathema Journal*. 2 (1), 57-66.

Irawati, R., Ekawati, Y. E., & Budiaawanti, S. (2020). Analisis Butir Ujian Akhir Semester Gasal Menggunakan Program Ambuso di SMA Negeri 1 Boyolali Tahun Ajaran 2019/2020. *Jurnal Materi dan Pembelajaran Fisika (JMPF)* 10(1) 2020 ISSN: 2089-6258.

Iskandar, A., & Rizal, M. (2017). Analisis Kualitas Soal di Perguruan Tinggi Berbasis Aplikasi Tap. *Jurnal Penelitian dan Evaluasi Pendidikan* 21 (2).

Istika N, Zamsir, & Prajono R. (2019). Kualitas Tes Ulangan Akhir Semester Mata Pelajaran
Anisa Putri, Yenita Roza, Maimunah
Analysis Of Mathematics Mid Test Questions For Class Vii Smp Education 21 Kulim
Anisa Putri, Yenita Roza, Maimunah

Analysis Of Mathematics Mid Test Questions For Class VII SMP Education 21 Kulim