The instrument for mathematics National Examination (UN) try out of Junior High School at Makassar City, South Sulawesi. Are there any revisions?

E Z Surni¹, H Retnawati²

¹Graduate Program of Mathematics Education, Yogyakarta State University, Depok, Sleman, Yogyakarta, Indonesia
²Department of Mathematics Education, Faculty of Mathematics and Natural Science, Yogyakarta State University, Depok, Sleman, Yogyakarta, Indonesia

*Corresponding author: ermazelfiana.2018@student.uny.ac.id

Abstract. Try out is an assessment form organized by the school and related parties before the implementation of national examination (UN) in Indonesia. By trying out, schools can conduct evaluations regarding the ability and readiness of students to face the UN. This research aims to determine the quality of the test instrument for UN mathematics try out of one of Junior High Schools at Makassar City, South Sulawesi. It can later be used as a revision basis. This is a descriptive-explorative type of quantitative research. The data is the test instrument and participant's answers, collected by documentation. The instrument is validated by the experts, and its quality is analyzed using the classical test theory approach. The results of the research indicate that the test instrument has Aiken index value of 0.854 which means that the test instrument has high validity for all items. The reliability coefficient of 0.82 means that the instrument has very good reliability. There are 11 difficult items, 14 items having a weak discrimination category and 4 items having ineffective distractor. The Items must be followed up by revising.

1. Introduction
Assessment is a component in the National Education Standard (SNP). This is explained in the government regulation of Indonesian [1]. Educational assessment is an assessment that aims to determine how well students are learning and is an integrated part of efforts to improve education [2]. The assessment will provide feedback to students, teachers, parents, policy makers, and the community about the effectiveness of educational services. According to Pellegrino et. al. [2], there are three purposes of the assessment, namely: supporting the learning process, measuring individual student achievement, and evaluating the program. This is supported by Brown [3] who argued that educational assessment refers to a set of methods and processes in which information or data about student learning is designed, collected, assessed, analyzed, and interpreted. This process aims to support decisions in the teaching process (example: what material needs to be re-taught by presenting different materials), in the learning process (for example: what material needs to be improved for students), in policy-making (example: what parts of the curriculum need to be developed by teachers). So it can be concluded that the educational assessment is a process of gathering and processing information related to the learning process of students and functions as a tool for evaluating the process and achievement of educational goals or curriculum.
According to *Permendikbud* about Educational Assessment Standards [4], it is explained that assessment is a process of gathering and processing information to measure the achievement of student learning outcomes consisting of three aspects namely attitude, cognitive, and skills. Basically, the assessment is carried out by three parties and has their respective objectives. The government is one of the parties that assesses education through a national examination (UN). The government assessment aims to assess the achievement of graduate competence in certain subjects nationally.

National Examination becomes one of the determinants of student graduation in an education unit for a certain period. In Indonesian, there are many ways that teachers and education units do to prepare the national exam. One of which is the try out test. Try out is a trial test conducted by schools or other related parties prior to the UN implementation. Noor and Aman [5] stated that try out is a form of trial exam for students, both to face semester exams and national exams. According to Arifa and Mardapi [6], a test instrument in try out test is adjusted to the characteristics of UN questions. Elfiza, Rusman, and Nasir [7] also stated that try out is divided into 2 important points. First, try out as an exercise and simulation for students. By practicing on questions similar to UN questions, students can get a description and get used to the form, type, and material of the questions to be tested at the UN. In addition, students can be trained to deal with procedures that are in the UN implementation. Thus they can situate themselves when faced with the real UN test situation. The second, try out as an evaluation for students, teachers, schools and the government. Through the try out test, students can see their weaknesses and material that they have not mastered so they intend to improve the quality of their learning. The teacher can find out the indicators that are the weaknesses of students so that they can take corrective steps in the teaching and learning process. For schools and the government, the try out results can be used as a basis for making policies in improving the quality of UN results.

The try out results are certainly expected to provide a valid description to find out how far the students' abilities and readiness in facing the national examination will be. These results can be a reference to teachers in predicting the percentage of students graduating in their respective schools. Therefore, in a try out test, it is necessary to use good quality questions so that the predictions and evaluations of teachers and schools can be accurate. According to Ebel and Frisbie [8], the quality of a good achievement test is closely related to the teaching implications and development of tests in the future. Therefore, teachers should assess the quality of the assessment tools. According to them, there are several reasons for the analysis of an instrument test. The low results obtained by students from a test might be due to the evaluation method not because of the lack of student learning intensity. Then the test must be evaluated, to determine whether the grades obtained by students are already in accordance with the expected goals. Teachers who have the right information about poor test items will have a good chance of improving classroom management and can clarify misconceptions obtained from test item data. The analysis and evaluation of tests are closely related to the development of teacher professionalism. By utilizing the data analysis, the teacher can have the ability to develop high-quality items.

To identify the quality of items, it is necessary to analyze the items. According to Ebel and Frisbie [8], there are some things that need to be considered in determining the quality of items, some of which are relevance, reliability, level of difficulty, and different power. Each analysis has its own criteria and is important to consider. Meanwhile, according to Retnavati [9], to obtain high-quality items, the items need to be analyzed theoretically that include content, construction, and language analysis. For multiple-choice questions, aspects of content can be seen from the effectiveness of the distractors. Besides that, it is also necessary to do an empirical analysis of items which includes an analysis of the difficulty index and discrimination index. In this research, the analysis used is focused on analyzing validity, reliability, difficulty index, discrimination index, and the effectiveness of the distractor.

2. **Method**

This research is descriptive-exploratory quantitative research. The purpose of this research is to determine the quality of the test instrument for UN mathematics try out. It can later be used as a
revision basis. The test participants consisted of 100 students in the third grade of the school. The data used in this study was the try-out test instrument in one of Junior High Schools at Makassar City, South Sulawesi. The test questions were chosen by the mathematics teacher at the school by taking from a collection of UN and USBN prediction questions. The test instrument was a multiple choice question consisting of 40 items. The research steps were initiated by documenting test questions and responses of students from related schools. After documenting, the researcher conducted the following analysis.

2.1 Validity
In the validity analysis, the test items were revalidated by 3 experts of mathematics education. The validation aimed to prove the content validity by considering the suitability of the items with the competency standards determined by the national education standard agency. Competency standards contained indicators that must be achieved by students on the 4 contents of mathematics material such as number, algebra, geometry and measurement, statistics and probability. The validity calculation used the Aiken index (V). Content validity criteria referred to Retnawati [9].

| Validity Index (V) | Criteria |
|-------------------|----------|
| V > 0.8           | High     |
| 0.4 < V ≤ 0.8    | Moderate |
| V ≤ 0.4           | Low      |

2.2 Reliability
Reliability analysis was conducted by looking internal consistency value on all items. According to Miller, Linn & Gronlund [10], The reliability index category should be based on the correlation coefficients in Table 2.

| Reliability   | Criteria      |
|---------------|---------------|
| 0.81 - 1.00   | Very Good     |
| 0.61 - 0.80   | Good          |
| 0.41 - 0.60   | Quite         |
| 0.21 - 0.40   | Poor          |
| 0.00 - 0.20   | Very Poor     |

2.3 Difficulty index
Analysis of difficulty index was conducted by looking at the percentage of students who answered the question items correctly. The following are the difficulty index criteria according to Retnawati [9].

| Difficulty Index (p) | Criteria |
|----------------------|----------|
| 0.00 ≤ p ≤ 0.29     | High     |
| 0.30 ≤ p ≤ 0.70     | Moderate |
| 0.71 ≤ p ≤ 1.00     | Low      |

2.4 Discrimination index
Analysis of discrimination index was conducted by looking at the value of point biserial on the correct answer and then classified according to criteria. The following are the difficulty index criteria according to Ebel & Frisbie [8] and Retnawati [9].
Table 4. Discrimination index criteria

| Discrimination Index (pbis) | Criteria    |
|----------------------------|-------------|
| 0.40 ≤ pbis                | Very good   |
| 0.30 ≤ pbis ≤ 0.39         | Good        |
| 0.20 ≤ pbis ≤ 0.29         | Not good    |
| pbis ≤ 0.19                | Poor        |

3. Findings

3.1 Validity and reliability
Based on the results of the content validity analysis, it was found that the Aiken index of the try out test instrument was 0.854. It means that the instrument has high validity for all items. For the reliability test, the internal consistency value was 0.82. It means that the test instrument provides stable and consistent measurement results, and small measurement errors [9].

3.2 Difficulty Index
The following are the results of the item difficulty index analysis.

Table 5. The result of difficulty index analysis

| Criteria   | Number of items                      |
|------------|-------------------------------------|
| Moderate   | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 18, 19, 20, 21, 22, 23, 28, 29, 31, 33, 35, 37, 38, 39, 40. |
| High       | 15, 16, 17, 24, 25, 26, 27, 30, 32, 34, 36 |

Based on the table, it can be interpreted that: 29 items or 72.5% of items have moderate criteria. 11 items or 27.5% of items have a difficult category and no items have an easy category.

3.3 Discrimination index analysis
The following are the results of the item discrimination index analysis.

Table 6. The Result of Discrimination Index Analysis

| Criteria       | Number of Items                      |
|----------------|-------------------------------------|
| Very good      | 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 18, 19, 20, 21, 22, 23, 28, 31, 38, 40 |
| Good           | 8, 35, 39                            |
| Not good       | 37                                   |
| Poor           | 15, 16, 17, 24, 25, 26, 27, 29, 30, 32, 33, 34, 36 |

Based on the table, it can be interpreted that 23 items or 57.5% items are very good, 3 items or 7.5% items are good, 1 item or 2.5% items is not good, and 13 items or 32.5% of items are poor.

3.4 The effectiveness of Distractor
In this study, there are 4 items that have ineffective distractors (≤ 5%), namely in numbers 17, 18, 21, and 24. In numbers 17 and 24, the ineffective distractor is A. In numbers 18 and 21, the ineffective distractor is D.

4. Discussion
In the following, the researcher will discuss two example items that need to be revised.
Number 15
If \( A = \{ x \mid \text{prime numbers} < 8 \} \) dan \( B = \{ x \mid -2 < x < 2, x \in \text{integers} \} \), the number of possible mapping from A set to B set is . . .
A. 27
B. 32
C. 54
D. 81

This item needs to be revised and should even be replaced with the following considerations: The context of the items in this item is actually open-ended and is not suitable if made in the form of multiple-choice questions. In calculating the number of maps students must map the elements of the two functions one by one. The calculation of mapping in this problem is very much and takes up a lot of time. Whereas in the study of several textbooks on material function, the researcher found that there was no material content or concepts in the book that computed a large number of mapping correlations. This is the reason why the questions are difficult for students, proven by the difficulty index of 0.17 where only 21 people are answered from 100 students. According to Sridadi [11], there are two possible problems that are not good or difficult for students. The first problem means ambiguity, especially in students who know a lot about the questions that are asked. The second, students have never learned things like those asked in the question. This item also has a negative discrimination index of -0.08 which means that this item is not able to distinguish high-ability students from low-ability students. A negative discrimination value indicates that the item is answered correctly by more students with low ability. The probable cause of the problem is because the problem is too difficult. According to Towns [12], questions that are too easy or too difficult will be difficult to distinguish between students who meet the measured learning objectives. Another drawback of this problem is the distractor B. Even though the percentage of students who chose answer B is more than 5%, the discrimination index of distractor B is positive. It shows that the distractor is more attractive to highly capable students. The fact is that the distractor B in this question was indeed chosen by 29 more highly skilled students compared to 10 students with low ability. According to Kartowagiran [13], good distractors will be chosen more by students with low ability. Highly capable students should be able to distinguish between distractors and answer keys. Meanwhile, according to Towns [12], in a multiple-choice question, the correct answer must have a positive discrimination index while all the distractors must be negative. If the distractor is positive, then the validity of the question must be questioned.

Number 24
Look at the following ABCD isosceles trapezium!

how many pairs of congruent triangles?
A. 7 pairs
B. 6 pairs
C. 5 pairs
D. 4 pairs
This question is considered difficult for students because it can only be answered correctly by 18 people out of 100 students. The discrimination index in this problem is negative. The stem of the question is actually clear and the answer key is also quite reasonable. The problem focus on this question lies with distractors A and B. A is considered to be ineffective because it is only chosen by 2 people or 2% percent of all students. According to Kartowagiran [13], the distractor is said to function if at least are selected by 5% of all students. In this question, there are 12 total triangles. Certainly, the A option doesn't make sense because to determine 7 pairs of congruent triangles, it takes 14 triangles. The A option can be made even simpler. For the B option, it was chosen by 60 students of 100 students. Sadly, this distractor is more plausible for highly skilled students. The number of 12 triangles is likely to be the cause of students directly determining B option. For follow-up revisions to this question, the teacher must pay attention to another aspect, namely the student's ability. The high level of difficulty or poor discrimination of questions can be caused by students' ability to absorb the material. This problem indicates that the teacher is not only required to make revisions of the question but also must increase the student's understanding of questions like this.

5. Conclusion
The results show that the content validity (Aiken index) is 0.845. It means that the test instrument is very valid for all items. The reliability coefficient of 0.82 means that the instrument has very good reliability. There are 11 items that are highly difficult, there are 14 items that have low discrimination index, and 4 items that have ineffective distractor. All results of the characteristic item analysis can be used as a revision basis in this try out test instrument. In this study, it is explained that there are several factors causing the characteristics of the items to be said bad. This information can be used by teachers in developing test instruments, especially test items and can be the basis for evaluating and improving the learning process. The researcher adds that another aspect that needs to be considered in writing questions or analyzing questions is the ability of students. The teacher is important to assess the suitability of the item characteristics with the cognitive level of students to ascertain whether the revision of the question really has to be done or it is precisely the student's ability that actually needs to be improved. In this study, the researcher uses the classical theory test model so that a challenge for future researchers is to try other test models such as Item Response Theory (IRT).

6. References
[1] R Indonesia 2005 Peraturan Pemerintah Republik Indonesia No. 19 Tahun 2005 Tentang Standar Nasional Pendidikan (Jakarta: Pemerintah Indonesia).
[2] J W Pellegrino, N Chudowsky, and R Glaser 2001 Knowing what Students Know (Washington, DC: National Academy Press).
[3] G T L Brown 2018 Assessment of student achievement (New York: Routledge).
[4] Kemendikbud 2016 Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia No. 23 Tahun 2016 Tentang Standar Penilaian Pendidikan (Jakarta: Kemendikbud).
[5] W N Noor and A Aman 2015 Evaluasi penyelenggaraan latihan ujian nasional bahasa inggris SMA swasta ex-RSBI J. Penelit. dan Eval. Pendidik. 19 pp 217–229.
[6] F N Arifa and D Mardapi 2014 Pengembangan perangkat tes try out UN SMP bidang studi matematika untuk pembentukan bank soal J. Eval. Pendidik. 2 pp 135–147.
[7] Y Elfiza, Rusman, and M Nasir 2016 Hubungan antara hasil uji kognitif try out ujian nasional (UN) dengan hasil ujian nasional (UN) mata pelajaran kimia SMA kota Banda Aceh tahun ajaran 2014/2015 Ilm. Mhs. Pendidik. Kim. 1 pp 35–42.
[8] R L Ebel and D A Frisbie 1991 Essentials of educational measurement. (Englewood Cliffs: Prentice-Hall International).
[9] H Retnawati 2016 Analisis kuantitatif instrumen penelitian (Yogyakarta: Parama Publishing).
[10] M D Miller, R L Linn, and N E Gronlund 2009 Measurement and assessment in teaching (New Jersey: Pearson Education Inc).
[11] Sridadi 2002 Analisis Butir Soal Pilihan Ganda J. Olahraga 8 pp 26–37.
[12] M H Towns 2014 Guide to developing high-quality, reliable, and valid multiple-choice assessments *J. Chem. Educ.* 91 pp 1426–1431.

[13] B Kartowagiran 2011 *Penulisan Butir Soal* (Yogyakarta: UNY).