ICSAT INTERNATIONAL PROCEEDING
ISBN: 978-623-7496-62-5
Vol, 11 Issue 4

The Development of Mathematics Final Assessment Instrument Test Second Semester in Grade VIII of Junior High School

Ruslan[1], Abdul Rahman[2], Husnul Khatimah Syam[3*]
[1] Mathematics Education, Universitas Negeri Makassar, Makassar, Indonesia
Email: ruslan@unm.ac.id
[2] Mathematics Education, Universitas Negeri Makassar, Makassar, Indonesia
Email: rahmanmallala@gmail.com
[3]Undergraduate Student of Mathematics Education, Universitas Negeri Makassar, Makassar, Indonesia
Email: husnulksyam@gmail.com

Abstract
This study aims to produce a final assessment test instrument second semester of Mathematics in grade VIII of Junior High School that is valid and reliable, as well as to know the validity, reliability, level of difficulty, discriminatory power, and distractor efficiency as the results of the development of Mathematics Second Semester final assessment instrument test in grade VIII of Junior High School. This research method is research and development (R&D) by adapting the seven development steps of Djemari Mardapi. The subjects in this study are two expert validators selected through the purposive sampling technique, and 182 students in grade VIII of UPT SPF SMP Negeri 26 Makassar were chosen through a random technique by agreement. The objects in this study are content validity, construct validity, reliability, difficulty level, discriminatory power, and distractor efficiency. The study result discovered that (1) the instrument test consists of 30 multiple choice items that meet the content validity aspect with the sum of an internal consistency coefficient is 1; (2) from the construct validity aspect, 24 items formed a factor, and all valid items are found in the 12 indicators that developed; (3) in terms of reliability aspect, the semester final assessment instrument test is reliable with the sum of the reliability coefficient is 0,74; (4) in terms of the level of difficulty, there are two items with difficult category, 18 items with medium category, and four items with the easy category; (5) in terms of discriminatory power, there are 19 items fill the excellent and medium category, so there is no need to be revised, two items with criteria needed to be revised, and three items with inadequate criteria; (6) in terms of distractor efficiency, there are 21 items have distractors that functioned well, and three items have distractors that functioned poorly. Thus, the semester final assessment instrument test developed in this study is feasible to measure students' Mathematics learning outcomes.

Keywords*: Research and development; Validity, Reliability; Level of difficulty.*

INTRODUCTION

The achievement of student learning outcomes is also known as evaluating. The evaluation aims to determine the effectiveness and efficiency of learning activities with the primary indicator of the success of learning activities to achieve the learning objectives that have been decreed (Suardipa & Primayana, 2020). Thus, the evaluation functions as part of the learning process because an assessment must always follow every learning activity to obtain a measurable picture of the extent to which students cope with the competency.

In preparing evaluation tools, teachers must adjust the assessment tools applied to essential competencies and indicators of competency achievement (Mauliandri et al., 2021). The preparation of evaluation tools not only aims to determine the achievement of student learning outcomes but as a teacher evaluation of whether the evaluation tool that has been arranged can carry out its function as a measuring tool for learning outcomes that have good quality or not (Fitrianawati, 2017).

Assessment of the achievement of student knowledge competence is an assessment in the cognitive domain consisting of levels of knowing, understanding, applying, analyzing, evaluating, and creating

(Prihatin, 2013). Assessment of the cognitive domain or knowledge of students can be done through written tests, oral tests, and assignments.

Teachers can determine the success of students' learning levels by using a test instrument expressed through a final score (Muthmainnah, & Purnamasari, 2019). One of the tools that can be applied to measure the understanding level of students in learning activities is using a test instrument. Teachers can determine the understanding level of students by giving a neat and qualified test. A test is said to be of good quality if the test has adequate validity and reliability (Satriani, 2019).

The need for item analysis in the learning process is by using tests that have been standardized or tested by the teacher (Rahmani et al., 2015). The standardized test is a test that has been reviewed and tested for the feasibility of the items. However, the teacher-made test is a test that is prepared directly by the teacher to evaluate the success of the student's teaching and learning process.

The test items prepared by the teacher must cover the entire curriculum, and the essential competencies must be completed. If the whole items are according to the curriculum content, then the items' validity is classified as high. If several items are not according to the curriculum content, then the items' validity is classified as low (Surapranata, 2005). Each item should have content validity, meaning that the measuring instrument contains the material that needs to be measured so that the suitability of the measuring instrument with the content that should be measured can be included in the writing of the items (Dwipayani, 2013).

The description shows the importance of the assessment aspect in improving the quality of the implementation of learning in schools because a good assessment produces good learning quality. A pre-research interview with one of the mathematics teachers at SMP Negeri 26 Makassar revealed that the teacher did not review the drafted items in the end-of-semester assessment. The test procedure and item analysis were not required, so the items tested on students did not meet the criteria for standardizing the instrument, as the results were not maximal. This non-maximization caused the learning outcomes for the students of grade VIII not to reflect the mastery of learning objectives comprehensively in Mathematics subjects. Besides that, the unknown quality of the items causes the teacher to be doubtful whether the test instrument that was made has prescribed its function accurately or not.

Based on the elucidation above, explain the need to produce a final assessment instrument test second semester of Mathematics in grade VIII of Junior High School that is valid and reliable, as well as to know the validity, reliability, level of difficulty, discriminatory power and distractor efficiency as the results of the development of Mathematics Second Semester final assessment instrument test in grade VIII of Junior High School.

METHOD

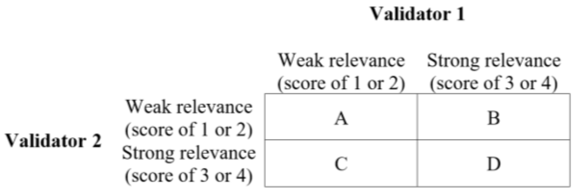
The type of this research method is research and development (R&D) by adapting the seven development steps of Djemari Mardapi, which are compiling test specifications, writing test items, reviewing test items, revising tests, implementing field trials, analyzing test items, and assemble instrument test (Mardapi, 2008). The development product is a final assessment instrument test second semester of Mathematics in grade VIII in the form of multiple choice.

This research is done from March to June 2022 at UPT SPF SMP Negeri 26 Makassar. The subjects in this study are two expert validators selected through the purposive sampling technique and 182 students in grade VIII of UPT SPF SMP Negeri 26 Makassar were chosen through a random technique by agreement, agreement (Agung, 2011) obtained between researchers, teachers, and students in grade VIII whose attend at the time of data collection, the student's presence when the data collection process is accomplished as if their arrival is considered randomly (Ruslan, 2010). The objects in this study are content validity, construct validity, reliability, level of difficulty, discriminatory power, distractor efficiency, and the development result of a final assessment instrument test second semester of Mathematics in grade VIII.

The data analysis was performed to know the content validity, construct validity, reliability, difficulty level, discriminatory power, and distractor efficiency of the developed test instrument.

2.1 Content Validity

The study of the content validity of the test instrument was performed to determine the suitability between the items and the lattice indicators on the test instrument. The expert validators can give a response to the assessment aspect by providing a score of 1 (irrelevant), 2 (slightly relevant), 3 (relevant), and 4 (very relevant).



		Validator 1	
		Weak relevance (score of 1 or 2)	Strong relevance (score of 3 or 4)
Validator 2	Weak relevance (score of 1 or 2)	A	B
	Strong relevance (score of 3 or 4)	C	D

Figure 1. The relevance category by two validators

Determine the whole content validity; it can be done through the calculation of the coefficient of internal consistency with the following formula:

$$\mathit{coefficient\ of\ internal\ consistency} = \frac{D}{A+B+C+D} \quad (1)$$

The determination of the degree of validity can be measured by using the agreement model of two validators to ensure the relevance of the items placed in cell D can reflect the valid agreement between validators. If the result of the validity coefficient is $\geq$ 0.75, thus it can be stated that the results of the measurements was made are valid (Ruslan, 2009).

2.2 Construct Validity

The instrument test's construct validity test was analyzed using the Confirmatory Factor Analysis (CFA) approach through the Maximum Likelihood Analysis (ML) method. Among the criteria required are: (1) the measurement results of Kaiser Meyer Olkin (KMO) Measure of Sampling Adequacy (MSA) > 0,50; (2) Bartlett's test with sig value 0,00 for further analysis; (3) Anti-Image Correlation (AIC) value with the measures criteria of sampling adequacy is ≥ 0,50 (Ruslan, 2006). The determination of homogeneous items is accomplished by taking into the factor loading ≥ 0,30 (Sappaile, 2006).

2.3 Reliability

To measure the test reliability value through multiple choice by using the Kuder Richardson 20 (KR-20) formula:

$$r_i = \frac{k}{(k-1)} \left\{ \frac{s_t^2 - \sum p_i q_i}{s_t^2} \right\} \qquad (2)$$

An instrument is said to be reliable if the Kuder-Richardson reliability coefficient is more than 0,70 ($r_i > 0{,}70$) (Yusup, 2018).

2.4 Level of Difficulty

The difficulty index for multiple choice type can be calculated by using the following formula:

$$P_i = \frac{n}{N} \qquad (3)$$

The results of the difficulty index calculation are interpreted in the following (Tilaar & Hasriyanti, 2019) categories:

Table 1. The Category of Difficulty Level Index Items

Category of level difficulty	Classification of level difficulty
Difficult	$0,00 \leq P \leq 0,30$
Medium	$0,31 \leq P \leq 0,70$
Easy	$0,71 \leq P \leq 1,00$

2.5 Distractor Efficiency

The distractor index can be calculated through the following formula:

$$IP = \frac{P}{(N-B)/(n-1)} \times 100\% \qquad (4)$$

The distractors that functioned properly were chosen by at least 5% of the students who took the test (Elvira & Hadi, 2016).

2.6 Discriminatory Power

The calculation of discriminatory power on each item can be done through the following formula:

$$DP = \frac{(WL-WH)}{n} \qquad (5)$$

The criteria for the enormity of the discriminatory coefficient are classified into four categories, which are good, intermediate (no need for revision), need for correction, and inadequate (Fitrianawati, 2017).

Table 2. The Discriminatory Power Category

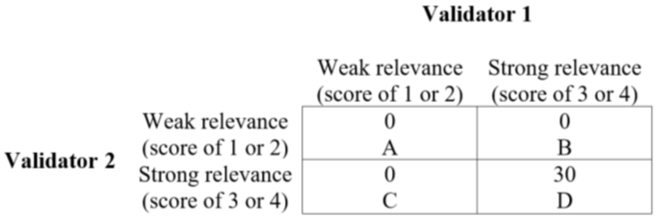
Discriminatory Power Category	Coefficient of Correlation
Good	$0,40 \leq DP \leq 1,00$
Intermediate (need for revision)	$0,30 \leq DP \leq 0,39$
Need for revision	$0,20 \leq DP \leq 0,29$
Inadequate	$-1,00 \leq DP \leq 0,19$

RESULT AND DISCUSSION

The steps include compiling test specifications, writing test items, reviewing test items, revising tests, implementing field trials, analyzing test items, and assembling instrument tests (Mardapi, 2008). The first step is compiling test specifications. The stage of preparing the test specifications includes determining the test's purpose, compiling the test content outline, and choosing the form of the test. The following stage is to write the test items. The number of items arranged in this study is adjusted to those listed in the test content outline. The coming stage is to examine the test items, which is done by asking for the help of two experts or validators. The aspect studied is the material aspect to assess the suitability between the items and the test content outline indicators on the instrument test. The validator study's results were then analyzed using the content validity with the Gregory testing model. After being reviewed by the validators, the next step is to improve the test based on the inputs and suggestions. The following stage is to conduct a test trial. Empirical data collection is done by testing the instrument test on students of grade VIII in UPT SPF SMP Negeri 26 Makassar. The tests that have been tested are then analyzed quantitatively to determine the construct validity, reliability, level of difficulty, distractor efficiency, and discriminatory power of the developed instrument test. After all, the items have been analyzed and corrected, the last step is to assemble the items into a complete test.

The development result of this study is that the second-semester final assessment test instrument formed of multiple choice contains 24 mathematics question items for the students of grade VIII. The development product is in the form of a final assessment test instrument in the second semester that has undergone two assessment stages: content validation by the expert validators and trial on students. The content validity involves two expert validators, and the test implementation involved 182 students of grade VIII in UPT SPF SMP Negeri 26 Makassar.

The final assessment test instrument second semester was developed through a process of reviewing items with content validity tests. Content validity reveals whether the statement items arranged in the test covered all the measured material (Budiastuti & Bandur, 2018). The content validity study on the test instrument was performed with the help of 2 experts to determine the suitability between the items and the test content outline indicators on the instrument. The results of the analysis of the validator are then analyzed using the content validity test with the Gregory validity test model.



		Validator 1	
		Weak relevance (score of 1 or 2)	Strong relevance (score of 3 or 4)
Validator 2	Weak relevance (score of 1 or 2)	0 A	0 B
	Strong relevance (score of 3 or 4)	0 C	30 D

Figure 2 The relevance category by two validators

Based on the content validity test, the internal consistency coefficient is 1 out of 30 questions that have been compiled. If the result of the validity coefficient is 0.75, it can be stated that the results of the measurements carried out are valid (Ruslan, 2009); because the result of the validity coefficient is 0.75, it can be stated that the result of the test instrument measurements that have been accomplished are declared valid (this indicates that the instrument items are valid).

The analysis of the end-of-semester assessment test instrument through the construct validity test was accomplished based on the result of the instrument test on students who were analyzed using the Confirmatory Factor Analysis (CFA) approach through the Maximum Likelihood (ML) analysis method. Based on the analysis that has been accomplished, the measurement result of the KMO MSA is 0.713 (KMO MSA > 0.50), and Bartlett's test shows a significant value of 0.000, so it is adequate for further analysis. Furthermore, the Anti-Image Correlation value, especially the correlation number marked a (diagonal direction from top left to bottom right) on 30 items, has an MSA value of 0.50, so it can be included in determining factors. Furthermore, with the Maximum Likelihood method, it is obtained that as many as 24 items indicate a load of each factor based on the indicator. Of the 30 items, it shows that there are 24 items from indicators 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 that have a factor loading value of 0.30. There are six items contained in indicator 4 (item 8 and item 9), indicator 7 (item 15), indicator 10 (item 23), indicator 11 (item 27), and indicator 12 (item 29), which shows the factor loading value < 0.3. Based on the result of processing the construct validity test, 24 items formed a factor, and all the valid items were contained in the 12 indicators developed.

Regarding reliability, the end-of-semester assessment test instrument has a coefficient of 0.74. the value of the reliability coefficient indicates that the items that have been used have stable or fixed characteristics. The interpretation of the test reliability coefficient is based on the opinion of Fraenkel et al. (Yusup, 2018), which states that an instrument is said to be reliable if the reliability coefficient value is more than 0.70.

The difficulty level of the items shows how easy or difficult the developed items are. Of the 24 items that have been declared valid and reliable, it was found that there were two items (8.3%) in the difficult category, 18 items (75%) in the medium category, and four items (16.7%) in the easy category. In preparing the items, the percentage of the difficulty level needs to be considered. According to Joesmani (Rahmani et al., 2015), a difficulty level between 25%-75% (enough category) is an adequate level of difficulty.

The discriminatory power aims to determine the ability of the items to distinguish students with high abilities and students with low abilities. The results of the analysis of the quality of the questions in terms of discriminating power, of the 24 items that were declared valid and reliable, there were 19 items (79.2%) that met the criteria of excellent and moderate, so that they did not need to be revised, there were two items (8.3%) with criteria need revision. The criteria need to be more suitable for as many as three items (12.5%).

The effectiveness of the distractor aims to determine the functioning of the available answers. The function of the distractor can be seen from the distribution of the answers of the participants who took the test. The distractor that worked well was chosen by 5% of the participants taking the test (Elvira & Hadi, 2016).

Based on the data from the distractor effectiveness analysis, it can be stated that of the 24 items declared valid and reliable, 21 items (87.5%) had distractors that functioned well. As many as three items (12.5%) had distractors that were not working well. Items 14, 19, and 25, whose distractors are not functioning properly, need to be corrected.

The quality of the developed items has met the valid and reliable criteria. The number of students who participated in the implementation of the test instrument development trial was 182 people. If from the number of respondents, the number has met the criteria because it is followed by a minimum of six times as many questions, as stated by Gable (Sappaile, 2006) that the number of respondents in the implementation of the trial is at least six to ten times the number of items to be analyzed. Thus, the semester final assessment instrument test developed in this study is feasible to measure students' Mathematics learning outcomes.

CONCLUSION

Based on the result of this study, the research on the development of the final semester assessment test instrument for mathematics subject of SMP grade VIII by adapting the seven development steps of Djemari Mardapi, which are compiling test specifications, writing test items, reviewing test items, revising test, implementing field trials, analyze test items and assemble instrument test. The instrument test consists of 30 multiple-choice items that meet the content validity aspect with the sum of an internal consistency coefficient is 1. From the construct validity aspect, 24 items formed a factor, and all valid items are found in the 12 indicators that developed. Regarding the reliability aspect, the semester final assessment instrument test is reliable, with the sum of the reliability coefficient being 0,74. In terms of the level of difficulty, there are two items (8,3%) in the difficult category, 18 items (75%) in the medium category, and four items (16,7%) in the easy category. In terms of discriminatory power, there are 19 items (79,2%) that fill the excellent and medium category, so there is no need to be revised; two items (8,3%) with criteria need to be revised, and three items (12,5%) with inadequate criteria. Regarding distractor efficiency, 21 items (87,5%) have distractors that functioned well, and three (12,5%) have distractors that functioned poorly. Thus, the semester final assessment instrument test developed in this study is feasible to measure students' Mathematics learning outcomes.

It is suggested that policymakers apply the test instrument in giving the odd semester final assessment test in the mathematics subjects of SMP class VIII. In addition, the test instrument developed was only up to the pilot stage for class VIII students at UPT SPF SMP Negeri 26 Makassar; therefore, to determine the effectiveness of the test instrument on a broader scope, it is recommended that enthusiasts apply it in other schools.

REFERENCES

Agung, I. G. N. (2011). *Manajemen penulisan skripsi, tesis, dan disertasi: kiat-kiat untuk mempersingkat waktu penulisan karya ilmiah yang bermutu*. PT RajaGrafindo Persada.

Budiastuti, D., & Bandur, A. (2018). *Validitas dan reliabilitas penelitian, dilengkapi analisis dengan NVIVO, SPSS, dan AMOS* (Vol. 1). Penerbit Mitra Wacana Media.

Dwipayani, S. (2013). Analisis validitas dan reliabilitas butir soal ulangan akhir semester bidang studi bahasa indonesia kelas X.D SMA N 1 terhadap pencapaian kompetensi. *Jurnal Pendidikan Bahasa Dan Sastra Indonesia Undiksha*, *1*(5). https://doi.org/https://doi.org/10.23887/jjpbs.v1i5.578

Elvira, M., & Hadi, S. (2016). Karakteristik butir soal ujian semester dan kemampuan siswa SMA di kabupaten muaro jambi. *Jurnal Evaluasi Pendidikan*, *4*(1), 58–68.

Fitrianawati, M. (2017). Peran analisis butir soal guna meningkatkan kualitas butir soal, kompetensi guru dan hasil belajar peserta didik. *Prosiding Seminar Nasional Dan Call for Papers Pendidikan 2017(PGSD UMS & HDPGSDI Wilayah Jawa)*. http://hdl.handle.net/11617/9117

Mauliandri, R., Maimunah, & Roza, Y. (2021). Kesesuaian alat evaluasi dengan indikator pencapaian kompetensi dan kompetensi dasar pada RPP matematika. *Jurnal Cendekia: Jurnal Pendidikan Matematika*, *5*(1), 803–811. https://doi.org/https://doi.org/10.31004/cendekia.v5i1.436

Mardapi, D. (2008). Teknik penyusunan instrumen tes dan nontes. In *Yogyakarta: Mitra Cendekia* (Vol. 127).

Muthmainnah, R. N., & Purnamasari, M. (2019). Analisis faktor penyebab peserta didik dengan IQ tinggi memperoleh hasil belajar matematika rendah. *FIBONACCI: Jurnal Pendidikan Matematika Dan Matematika*, *5*(1), 81–86. https://doi.org/https://doi.org/10.24853/fbc.5.1.81-86

Prihatin, S. (2013). *Model penilaian pencapaian kompetensi peserta didik sekolah menengah pertama*. Kementrian Pendidikan dan Kebudayaan.

Rahmani, M., Ningsih, K., & Nurdini, A. (2015). Analisis kualitas butir soal buatan guru biologi kelas X SMA Negeri 1 Tanah Pinoh. *Jurnal Pendidikan Dan Pembelajaran Khatulistiwa*, *4*(2). https://doi.org/http://dx.doi.org/10.26418/jppk.v4i2.8970

Ruslan. (2010). *Penilaian kinerja dosen berdasarkan kepuasan mahasiswa dan pengaruhnya terhadap perilaku pascakuliah (studi di FMIPA universitas negeri makassar)*. Pustaka Yaspindo.

Ruslan. (2009). Validitas isi. *Buletin Pa'biritta LPMP Sulawesi Selatan*, *IV*(10), 18–19.

Ruslan. (2006). Aplikasi analisis faktor dalam uji-validitas instrumen penelitian. *Seminar Nasional Statistika IKAPSTAT ITS Dan Jurusan Matematika FMIPA UNM*, 1–11.

Satriani. (2019). *Analisis soal ulangan akhir semester mata pelajaran matematika kelas VI SD di kecamatan malili kabupaten luwu timur* [Universitas Negeri Makassar]. http://eprints.unm.ac.id/14085/

Sappaile, B. I. (2006). Dimensi dan reliabilitas suatu instrumen dengan menggunakan rotasi varimax pada analisis faktor eksploratori. *Jurnal Pendidikan Dan Kebudayaan Tahun Ke*, *12*(060), 1–15.

Suardipa, I. P., & Primayana, K. H. (2020). Peran desain evaluasi pembelajaran untuk meningkatkan kualitas pembelajaran. *Widyacarya: Jurnal Pendidikan, Agama Dan Budaya*, *4*(2), 88–100. https://doi.org/https://doi.org/10.55115/widyacarya.v4i2.796

Surapranata, S. (2005). *Analisis, validitas, reliabilitas, dan interpretasi hasil tes* (2nd ed.). Remaja Rosdakarya.

Tilaar, A. L. F., & Hasriyanti, H. (2019). Analisis butir soal semester ganjil mata pelajaran matematika pada sekolah menengah pertama. *Jurnal Pengukuran Psikologi Dan Pendidikan Indonesia (JP3I)*, *8*(1), 57–68. https://doi.org/10.15408/jp3i.v8i1.13068.

Yusup, F. (2018). Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, *7*(1), 17–23. https://doi.org/http://dx.doi.org/10.18592/tarbiyah.v7i1.2100