JERE 8 (1) (2019) 73 - 83

Journal of Educational Research and Evaluation

http://journal.unnes.ac.id/sju/index.php/jere

Development of Test Instrument TIMSS Model to Measure Mathematical Ability VIII Grade Students

Nur Romdlon Maslahul Adi ✉, **Kartono, Endang Susilaningsih**

Universitas Negeri Semarang, Indonesia

Article Info

Article history:
Received 25 December 2018
Approved 08 January 2019
Published 10 March 2019

Keywords:
TIMSS, R&D, mathematical ability

Abstract

The teachers are still lack found practice tests that have characteristics such as the Trend in International Mathematics and Science Study (TIMSS) in Junior High School mathematics textbooks. This study aims to develop a TIMSS test instrument model that refers to the cognitive domain and the content domain for class VIII students. The respondents of this study were 5 people in the expert test, 10 people in the one to one test and 53 people in the small-scale test. This research is development research conducted following the development model of Borg & Gall. This step is only taken until the ninth step with a few modifications. Modifications are made by adding a one to one test copy from Tessmer. In content validity, as many as 5 items out of 35 items developed are not yet valid with a coefficient range of 0.6-0.79. Items were then revised to be tested in the small-scale test stage. One to one test results showed the instrument readability level was 81%. The validity test of the criteria shows a coefficient of 0.84 which means that the product developed is comparable to the original product. Interrater reliability was 0.661 and Alpha Chronbach reliability testing was a small scale at 0.854. The results showed that the instrument developed was valid in terms of content and criteria as well as reliable for use. The benefit of this research is that the developed test instrument can be used as a tool to measure the mathematics abilities of eighth-grade students.

✉Correspondence Address:
Kampus PAscasrjana Unnes Jl kelud utara 3, Sampangan, Semarang, Indonesia'
E-mail: romdlon.adi@gmail.com

p-ISSN 2252-6420
e-ISSN 2503-1732

INTRODUCTION

Minister of Education and Culture Regulation (Permendikbud) Number 21 of 2016 concerning Basic and Secondary Education Content Standards states that competencies in mathematics learning at the Junior High School / Madrasah Tsanawiyah (MTs) level are expected in each material one of which is to show a logical attitude, critical, analytical, meticulous and thorough, responsible, responsive, and does not give up easily in solving problems. Students are also expected to have confidence, group interaction, and an interest in mathematics. The comparison of students' abilities with the basic competencies expected above can be seen from the results of research conducted by the Trend in International Mathematics and Science Study (TIMSS) that Indonesia participated in.

The TIMSS results to date still place the ability of Indonesian students in mathematics and science in the lower class. The average ability of Indonesian students is still below the average ability of students their age in the world. In 2011, Indonesia ranked 45th out of 50 countries (Mullis, Martin, Foy, & Arora, 2012). The percentage of reasoning is the weakest result ability of Indonesian students (Vendiagrys, Junaedi, & Masrukan, 2015, p. 35). The TIMSS study students' assessment of mathematical abilities based on the content domain and cognitive domain. The content domain includes numbers, algebra, geometry, and data and opportunities. The cognitive domain tests students' mathematical abilities in aspects of knowledge, application, and reasoning (Ina V.S. Mullis and Michael O. Martin, 2017, p. 7).

The results of Indonesia's achievements in the low TIMSS study were caused by several factors. One contributing factor is because students in Indonesia are not trained enough to solve contextual questions, demand reasoning, argumentation and creativity which are characteristic of TIMSS questions. Whereas the support of teachers in the classroom in the form of providing materials that hone the ability to think at a high level has a positive effect on the ability of students (Arends, Winnaar, & Mosimege, 2017, p. 2). The used of learning models that practice higher-order thinking skills can improve the ability of analysis, evaluation, and the ability to create sustainable learning processes in students (Diputera, Setyowati, & Susilaningsih, 2018, p. 63). The process of thinking in problem-solving requires the attention of the teacher to assist students in developing problem-solving abilities in both the real-world context and the mathematical context (Ulya, Kartono, & Retnoningsih, 2014, p.1).

Teachers were still having trouble finding practice questions that have characteristics such as TIMSS problems in junior high school mathematics textbooks. Books are many students wrestle with in daily learning (Wardhani, 2011, p. 61). Therefore, it is necessary to develop a valid and reliable TIMSS model test instrument so that it can measure according to what is being measured and be consistent when used.

Several TIMSS model test instrument developments have been carried out several times. Vebrian, Hartono, & Darmawijoyo, (2016) developed the TIMSS model problem for the Number domain. Hazlita, Zulkardi, & Darmawijoyo (2014) took the cognitive domain of Reasoning as a focus on the development of 10 TIMSS questions. Tri Wahyudi, Zulkardi (2016) focused on the question of Reasoning in its development. Rizta, Zulkardi, & Hartono (2013) only developed TIMSS multiple-choice questions. Researchers found many studies that developed the TIMSS question model. Development is still limited by MCQs or relatively small numbers. Some of the development that was carried out also only took a part of the TIMSS domain aspect. Based on the explanation above, the TIMSS model test instrument to measure the reasoning of students of Mathematics VIII grade in junior high school needs to be arranged in a more complete form,

namely multiple choice and structured responses (matched, true-false, stuffed).

METHOD

The research design used in this research is development research. The development research procedure carried out follows the Borg & Gall development model (Gall, Gall, & Borg, 2003, p. 569) which includes ten steps of the research strategy and implementation. The ten steps are (1) preliminary studies, (2) planning, (3) development of initial models or products, (4) review of hypothetical models, (5) revisions, (6) small-scale tests, (7) revision of experimental results, (8) large-scale testing, (9) revision of the final model, (10) dissemination.

This step is only taken until the ninth step with a few modifications. Modifications are made by adding one to one test copy from Tessmer (2013, p. 15). A one-to-one trial was conducted to determine the level of readability of students towards instruction in long questions and tests (Pulungan, 2014, p. 76). This article will explain the process of developing TIMSS model questions into the small-scale test phase.

Content validity was assessed by 5 professional assessors. Reliability was assessed by the Interclass Correlation Coefficient (ICC) because the assessor was more than 3. Reliability was analyzed using SPSS 24. The criterion validity test was carried out at the small-scale test stage. Criteria validity test is done by comparing the results of working on TIMSS model questions made with the results of the original TIMSS questions. Comparisons are made by paying attention to the same content domain and the cognitive domain. Analysis of the different levels of difficulty and strength was carried out at the small-scale test stage with the research subjects eighth graders of junior high school in Kudus.

RESULT AND DISCUSSION

Previous Study

The previous study in the form of a needs analysis was carried out by conducting interviews with mathematics teachers in class VIII. Interviews were conducted to find out the implementation of mathematics learning and the process of making questions that were used to assess students' mathematical abilities. Previous studies were also carried out with document studies to see what the TIMSS problem looks like and the domains that exist in the TIMSS model problem.

Planning of the TIMSS model test instrument that will be developed has been carried out. Planning is done by making a test instrument lattice the TIMSS model developed. Researchers prepared as many as 35 items that have been divided according to the proportion of the cognitive domain and the content domain. The proportions of the TIMSS model items developed are outlined in Table 1.

Table 1. The Proportions of The TIMSS Model Items

Domain		*Knowing* (35%)	*Applying* (40%)	*Reasoning* (25%)	Total
Number	(30%)	4	4	3	11
Aljabar	(30%)	3	4	3	10
Geometry	(20%)	2	3	2	7
Data & Chance	(20%)	2	3	2	7
Total		11	14	10	35

Content Validity

The TIMSS model test instrument developed for grade VIII students was tested for

content validity by several experts. The content validity test was conducted by 5 experts to see the suitability of the material, construction, and language of the instrument being developed. The instrument validators were 3 academics and 2 practitioners teaching in junior high schools.

Expert validators assessment the quality of TIMSS model questions created and their suitability with the content domain and cognitive domain. The results of the test instrument expert judgment are then calculated with the Aiken V formula that has been designed in Microsoft Excel calculations.

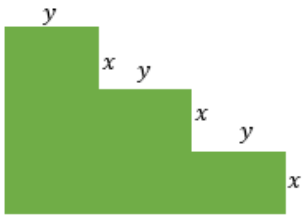
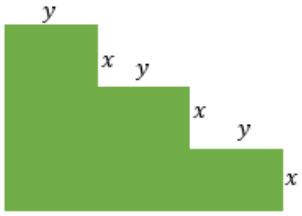
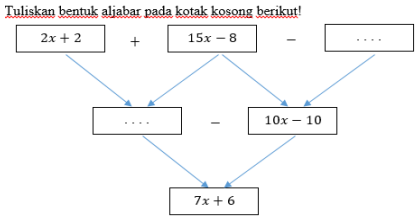
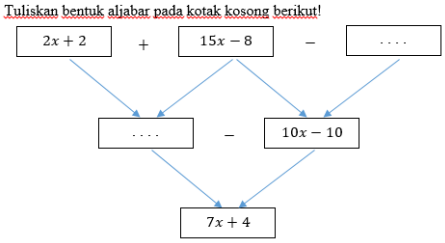
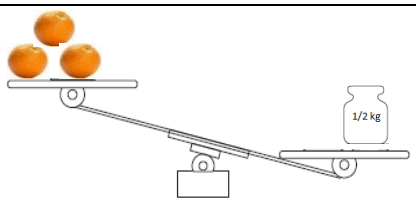
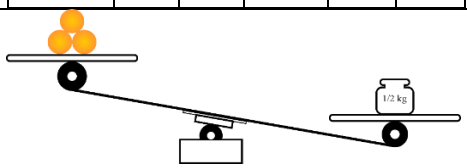
Table 2. The Coefficient Value of Expert Approval

Indeks Aiken's V	Result	Number of Item
0.8-1.0	Valid	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 17, 18, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35
0.6-0.79	Invalid	15, 19, 20, 21, 25

The expert approval coefficients obtained were then compared with the Aiken's coefficient validity table version. The item is said to be adequate or valid if the coefficient value of expert approval> V. Table V Aiken shows that the value of V for the 5 evaluators and 5 choice scales is validity> 0.80 (Aiken, 1985). Conversely, the coefficient of the expert agreement is said to be inadequate or not 0.80 for a 0.04 error chance. Data for calculating content validity with the Aiken's V formula is shown in Table 2.

Based on Table 2, the number of valid items is 30 items. While items that are invalid are 5 items. Items that are invalid are not discarded but are revised again to be retested at a later stage because they still have a high Aiken V index. The researcher then revised the items that were invalid to be tested again at the one to one test stage. The revision was carried out by looking at the results of the validity test with Aiken's V and qualitative data obtained by the researchers from the experts' recommendations. The researcher revised 5 items that were invalid. The changes made were related to the suggestions given by the experts. The revisions made by the researchers are shown in Table 3.

Table 3. The Revision on Invalid Item

Item	Before Revision	Revision
15	The proportion size of picture is not suitable. 	The proportion size of picture is suitable 
19	The written mistaken of Aljabar form: $7x + 6$ 	$7x + 6$ turned into $7x + 4$ 
20	x | -1 | | 1 | $f(x)$ | | 5 | 2 | -1	x | -1 | | 1 | | $f(x)$ | | 5 | 2 | | -1 (x, y) | | | (1,2) | |
21		
25	The picture is not clear	The picture is clear

Interrater Reliability

After doing the content validity from the experts, the results of the existing research were then calculated the level of agreement between the five experts using a consistency reliability test between assessors using the Intraclass Correlation Coefficient (ICC) analysis. ICC analysis was carried out with the help of SPSS 24.

ICC's analytic results are shown in Table 4. ICC analysis results obtained r_{xy} reliability of 0.661. According to Rusilowati (2014, p. 29), the reliability criteria were considered to be in the range of values of 0.4 ≤ r <0.6. Reliability is considered high if it is in the range of 0.6 ≤ r <0.8. Based on this classification, the reliability of 0.661 is included as high reliability.

Table 4. Intraclass Correlation Coefficient

		95% Confidence Interval		F Test with True Value 0			
	Intraclass Correlation[a]	Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.280[b]	.138	.461	2.949	34	136	.000
Average Measures	.661[c]	.444	.811	2.949	34	136	.000

One to One Test

After got the results of the content validity and interrater reliability, the researchers tested 35 items in the one-to-one trial phase. One to one trial is given to 10 students who are chosen with a variety of low, medium, and high abilities. The determination of students who become one to one test respondents is based on teacher recommendations based on ability levels. The researcher then confirms these classifications with the previous mathematical values document owned by the teacher.

Data from one to one trial was obtained by providing a reability questionnaire containing 10 questions. The recapitulation of data on the readability questionnaire by students is shown in Table 5.

Table 5. Results of Student Readability Questionnaire

No	Questionnaire	Result (%)		Number of Item
		Yes	No	
1.	Is there a word or expression clearly illegible ?	0	100	
2.	Is there too big letters?	0	100	
3.	Is there unclear working instruction?	40	60	15
4.	Is there a confusing question?	80	20	11, 16, 25, 32
5.	Is there a term understandable ?	0	100	
6.	Is there a difficult formula to understand?	20	80	27
7.	Is there a question you did not in accordance with a description of ?	0	100	
8.	Is there a question which the answer more than one?	50	50	22
9.	Is there unreadable image?	0	100	
10.	Is there a chart , table , or diagram illegible clear ?	0	100	
	Average	19	81	

The results of data processing in Table 5 show the instrument readability level is 81%. Data on readability aspects were also taken with a brief interview technique with students after completing the questionnaire. The results of the discussion obtained qualitative data as follows: 1) there are some questions that students have difficulty understanding the questions, 2) there are questions that are difficult to work on, but there are also problems that are easy to work on, 3) there are questions that are not in accordance with those in the book, 4) students are not accustomed to working on this type of problem.

Question number 3 asks for unclear working instructions. As many as 4 out of 10 students who took the one to one test pointed to item number 15 in the form of matched questions as unclear items. Students are not

accustomed to finding matched questions so there is a bit of confusion.

Question number 4 about whether there are confusing questions, 80% of respondents answered that there were confusing questions. The items that were considered confusing by respondents varied, namely item 11, 16, 25, and 32. The researcher conducted interviews with respondents to find out which parts of the questions were confusing. Item number 11 raised confusion in the order to check because students were not accustomed to working with questions like item number 11 in the Figure 1.

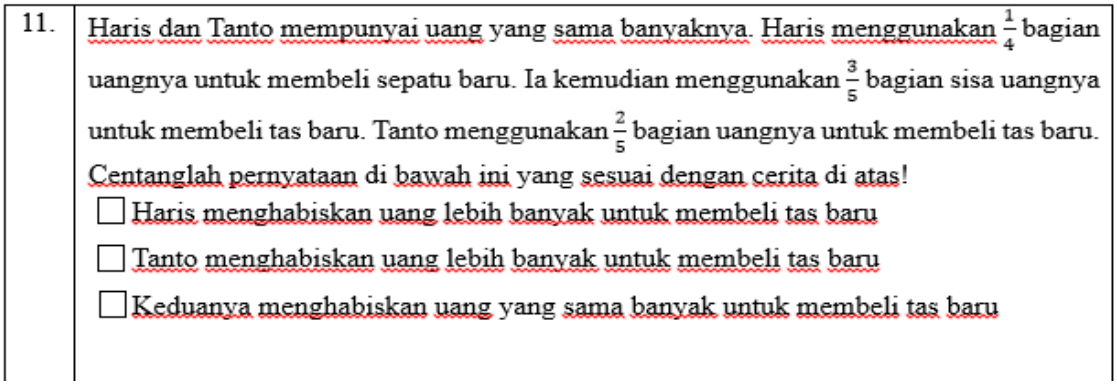


Figure 1. Item Number 11

Item number 16 raised confusion for students on the question "what is the longest side length of the triangle above?". The question was then replaced by researchers to "what is the longest side value of the triangle?". Item number 25 raises student's confusion about the purpose of the number of angles, that is, between the number of angles or the number of degrees from all angles. The researcher then marks the degree in the answer section so students know what kind of answer the question maker is asking.

Question number 6 asks about a difficult formula to understand. In that question, 2 of 10 students wrote that item 27 number was difficult to understand. The difficulty of students is more to students forgetting the formula or method used to work on the problem model to determine the angle in a triangle. The researcher decides not to revise the item.

Question number 8 asks about any questions with more than one answer. In that question, 5 of 10 students who took the one to one test answered item 22 with more than one answer. After the researcher reviewed, it turned out that there was an error in making the answer option so that there were two correct answers to the option. Changes to these items are shown in Figure 2.

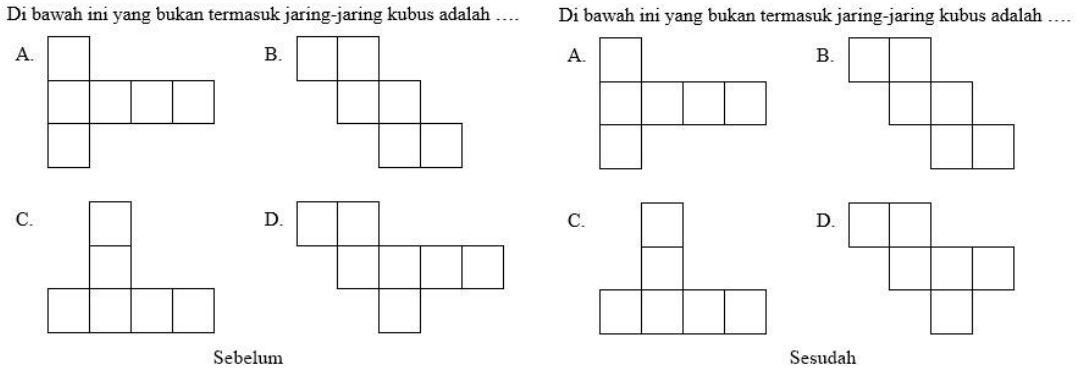


Figure 2. Comparison of Change The Revised

Criteria Validity

Criteria validity test was carried out at the small-scale test stage. Criteria validity test is done by comparing the results of working on TIMSS model questions made with the results of the original TIMSS questions. Intake of grades is carried out simultaneously during small scale tests. Six questions from 35 questions were selected equally based on the content domain and cognitive domain about the TIMSS model. The selected questions are then compared with the original TIMSS questions that have the same domain as in Table 6.

Table 6. Number of Item That Analyzed The Criteria Validity

Domain Konten	Domain Kognitif	TIMSS Model Questions	The Original TIMSS Questions
Number	Knowing	1	36
Data & Chance	Knowing	29	37
Algebra	Knowing	13	38
Geometry	Knowing	22	39
Geometry	Reasoning	27	40
Algebra	Applying	17	41

Researchers took 6 original TIMSS items from the 2011 and 2015 TIMSS item collections. Four items were taken from four different domains with the level of cognitive domain knowledge. Then the researchers took algebraic material items with the level of application and geometry problems at the level of reasoning. The validity test results show that the value of r_{xy} is 0.84, indicating that the developed TIMSS model has comparative measurement capabilities with the original TIMSS problem.

Items Characteristics

Analysis of the characteristics of the TIMSS model test instrument to measure the mathematical abilities of eighth-grade students consists of different strengths and difficulty levels. Different strengths of questions are used to analyze differences in abilities between individual students (Rahayu & Purnomo, 2014, p. 41). The different power classifications according to Croker and Algina in Rusilowati (2014, p. 38) are $0.00 \leq D < 0.20$ questions discarded, $0.20 \leq D < 0.30$ questions corrected, $0.30 \leq D < 0$, 40 questions were accepted but need to be corrected, $0.40 \leq D < 1.00$ questions accepted. Data on the results of different power analyzes are presented in Table 7.

Table 7. Consists of Different Strengths

D	Result	Number of Item
$0.00 \leq D < 0.20$	items removed	2, 5, 8, 9, 12, 14, 18, 25, 26, 28, 30, 33, 34
$0.20 \leq D < 0.30$	items repaired	1, 3, 4, 11, 15, 16, 19, 21, 22, 35
$0.30 \leq D < 0.40$	items received with repairs	7, 10, 24, 31
$0.40 \leq D < 1.00$	items received	6, 13, 17, 20, 23, 27, 29, 32

Table 8. The Level of Difficulty

TK	Result	Number of Item
$0.00 \leq$ TK ≤ 0.30	Soal sukar	7, 16, 24, 25, 27, 29, 30, 31, 32, 34
$0.30 < TK \leq 0.70$	soal sedang	2, 6, 9, 10, 11, 12, 13, 15, 17, 19, 20, 21, 23, 26, 28, 33, 35
$0.70 < TK \leq 1.00$	soal mudah	1, 3, 4, 5, 8, 14, 18, 22

Different power analysis results show that of the 35 items tested, as many as 8 items were received, 4 items were received with repairs, 10 items were repaired, and 13 items were removed. Researchers make improvements by analyzing various strengths and discarding items that cannot distinguish students' abilities. Researchers discard item numbers 2, 5, 8, 9, 12, 14, 18, 25, 26, 28, 30, 33, and 34 which have low power levels so it is not recommended for use. Researchers also make improvements to items with criteria that are accepted with a slight improvement and items with criteria must be corrected.

The characteristic of the item sought besides the different strengths is the level of difficulty of the questions. The level of difficulty of an item is an opportunity to answer the problem correctly at a certain ability level (Undorf, Erdfelder, Undorf, & Erdfelder, 2013). The results of the analysis of items on a small scale test showed that the difficulty level of 35 items varied, ranging from easy, medium, to difficult. The difficulty level classification of the questions is $0.00 \leq TK \leq 0.30$ difficult questions, $0.30 \leq TK \leq 0.70$ moderate questions, and $0.70 \leq TK \leq 1.00$ easy questions (Rusilowati, 2014, p. 35). Difficulty index of 0.0 indicates that the problem is too difficult, while an index of 1.0 indicates that the problem is too easy (Solichin, 2017, p. 196). The results of item analysis on small scale tests can be seen in Table 8.

The results of the analysis of the difficulty level showed that of the 35 questions, there were 8 items with easy difficulty levels, 19 items with moderate criteria, and 8 items with difficult criteria. The level of difficulty that is easy or difficult is not eliminated because of the nature of the test used to measure students' mathematical abilities so that the items needed with criteria ranging from easy, medium, to difficult. This criterion corresponds to the level of the cognitive domain, namely knowing, applying, and reasoning.

Analysis of item characteristics based on Classical Test Theory eliminates items that have a low power differential. The results of the analysis of item characteristics based on the Classical Test theory which links the results of different power levels and difficulty levels are explained through Table 9.

Table 9. Item Characteristics

Item Characteristics	No. Butir
Difficult	7, 16, 24, 27, 29, 31, 32
Medium	6, 10, 11, 13, 15, 17, 19, 20, 21, 23, 35
Easy	1, 3, 4, 22

Reliability

Small-scale reliability analysis was performed using the SPSS application. The steps taken are entering data then click Analysis,

Scale, Reliability Analysis menu then select an Alpha model. The reliability results are shown in Table 10.

Tabel 10. Reliability Statistics

Cronbach's Alpha	N of Items
.854	35

The alpha value is seen based on the results of the Alpha Cronbach Reliability Statistics output. Cronbach's Alpha small scale reliability test results showed a value of 0.854. Rusilowati (Rusilowati, 2014, p. 29) classifies the criteria for reliability as very low (r <0.2), low (0.2 ≤r <0.4), moderate (0.4 ≤r <0.6), high (0.6 ≤r <0.8), and very high (0.8 <r <1.0). Based on this classification, the reliability value of 0.854 is included in the high category. The higher the reliability coefficient, the more consistent the instrument is if it is used repeatedly (Khumaedi, 2012, p. 29).

CONCLUSION

Based on the results of research and discussion, it can be concluded that the TIMSS model test instrument developed has been valid in terms of content and has a good difference in power. From 35 items was developed, 22 items were found that could distinguish students' abilities. 13 items were eliminated because they could not distinguish students' abilities.

The validity of the developed TIMSS model test instrument criteria has a high-value coefficient, which is 0.84. The reliability of the developed test instrument is also quite good. Interrater reliability of 0.661 changed to 0.855 at the small-scale test stage so that it was categorized as very high.

ACKNOWLEDGMENT

Acknowledgments the researchers convey to those who have helped during the research process, including: 1) Dr. Dafid Slamet Setiana, M.Pd. as an expert validator, 2) Musyiana S, S.Pd and Adista Zelmy V., S.Pd. as a teacher at SMP N 1 Dawe Kudus as well as an expert validator, and 3) SMP N 1 Dawe Kudus as a place of research that provides researchers the opportunity to take research data.

REFERENCES

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. Educational and Psychological Measurement, 45(1), 131–142.

Arends, F., Winnaar, L., & Mosimege, M. (2017). Teacher Classroom Practices and Mathematics Performance in South African Schools: A Reflection on TIMSS 2011. South African Journal of Education, 37(3), 1–11. https://doi.org/10.15700/saje.v37n3a1362

Diputera, A. M., Setyowati, D. L., & Susilaningsih, E. (2018). Higher-Order Thinking Skills of Junior High School Students. The Online Journal of New Horizons in Education, 8(3), 61–67.

Gall, M. G., Gall, J. P., & Borg, W. R. (2003). Educational Research: An Introduction, 7th Edition. Boston: Longman Publishing.

Hazlita, S., Zulkardi, & Darmawijoyo. (2014). Pengembangan Soal Penalaran Model TIMSS Konteks Sumatera Selatan di Kelas IX SMP. Jurnal Kreano, 5(November), 170–179.

Ina V.S. Mullis and Michael O. Martin. (2017). TIMSS 2019 Assessment Frameworks.

Khumaedi, M. (2012). Reliabilitas Instrumen Penelitian Pendidikan. Jurnal Pendidikan Teknik Mesin, 12(1), 25–30.

Mullis, I. V. S., Martin, M. O., Foy, P., & Arora, A. (2012). TIMSS 2011 International Result in Mathematics. Boston: TIMSS & PIRLS International Study Center.

Pulungan, D. A. (2014). Pengembangan Instrumen Tes Literasi Matematika Model PISA. Journal of Educational Research and Evaluation, 3(2), 2–6.

Rahayu, T. D., & Purnomo, B. H. (2014). Analisis Tingkat Kesukaran dan Daya Beda Pada Soal Ujian Tengah Semester Ganjil Bentuk Pilihan Ganda Mata Pelajaran Ekonomi Kelas X di SMA Negeri 5 Jember Tahun Ajaran 2012-2013 (The Analysis of Difficulties and Distinguishing Power on The Middle Test wi. Jurnal Edukasi UNEJ, 1(1), 39–43.

Rizta, A., Zulkardi, Z., & Hartono, Y. (2013). Pengembangan Soal Penalaran Model TIMSS Matematika SMP. Jurnal Penelitian Dan Evaluasi Pendidikan, 17(2), 230–240. Retrieved from https://journal.uny.ac.id/index.php/jpep/article/view/1697

Rusilowati, A. (2014). Pengembangan Instrumen Penilaian. Semarang: Unnes Press.

Solichin, M. (2017). Analisis Daya Beda Soal, Taraf Kesukaran, Validitas Butir Tes, Interpretasi Hasil Tes dan Validitas Ramalan dalam Evaluasi Pendidikan. Dirāsāt: Jurnal Manajemen Dan Pendidikan Islam, 2(2), 192–213.

Tessmer, M. (2013). Planning and Conducting Formative Evaluations. Routledge.

Tri Wahyudi, Zulkardi, D. (2016). Pengembangan Soal Penalaran Tipe TIMSS Menggunakan Konteks Budaya Lampung. Jurnal Didaktik Matematika, 3(1).

Ulya, H., Kartono, K., & Retnoningsih, A. (2014). Analysis of Mathematics Problem Solving Ability of Junior High School Students Viewed from Students' Cognitif Style. Journal of Education and Practice, 2(10).

Undorf, M., Erdfelder, E., Undorf, M., & Erdfelder, E. (2013). Separation of Encoding Fluency and Item Difficulty Effects on Judgements of Learning. Quarterly Journal of Experimental Psychology, 66(10), 2060–2072. https://doi.org/10.1080/17470218.2013.777751

Vebrian, R., Hartono, Y., & Darmawijoyo. (2016). Pengembangan Soal Matematika Tipe TIMSS Menggunakan Konteks Kerajaan Sriwijaya di SMP. Jurnal Didaktik Matematika, 3(2), 96–105.

Vendiagrys, L., Junaedi, I., & Masrukan. (2015). Analisis Kemampuan Pemecahan Masalah Matematika Soal Setipe TIMSS Berdasarkan Gaya Kognitif Siswa Pada Pembelajaran Model Problem Based Learning, 4(1), 34–41.

Wardhani, S. (2011). Instrumen Penilaian Hasil belajar matematika SMP: Belajar dari PISA dan TIMSS. Kementerian Pendidikan Nasional.