AlphaMath
Journal of Mathematics Education, 9(2) November 2023

Analysis of Pre-Olympic Middle School Mathematics Test Instruments Based on Item Response Theory

Dwi Cahyani Nur Apriyani*, Hari Purnomo Susanto, Taufik Hidayat
Mathematics Education Department, STKIP PGRI Pacitan East Java, Indonesia
*yaalatiif09@gmail.com

ABSTRACT

Junior high school of mathematics Olympiad is one of the routine activities every year in Indonesia. Analysis of the items on the Mathematical Olympiad test instrument has been carried out by several researchers using the concept of classical test theory. The theory has many weaknesses when compared to the concept Item Response theory (IRT). The purpose of this article is to describe the analysis of pre-Olympiad mathematics test instrument for junior high school students in Pacitan East Java based on the IRT. This analysis was carried out using a quantitative descriptive method. IRT-based item analysis was performed using R software with the package irtawsi. The results of the analysis with the package show that the instruments used fit to 2PL model. From the 15 items used, there were two items that did not fit, namely items 4 and 9. Furthermore, from the 13 test items that were fit, there were 3 items that did not meet the quality of the discriminant parameter used. These three items are 1, 6 and 13 items, these three items must be dropped because they can provide biased information when used to estimate students' math olympiad abilities. Based on this information it can be concluded that there are 10 test items that have quality that meets the qualification parameters of difficulty and discriminant with a standard error of 0.25.

Keywords: IRT, Item Parameters, Item Quality, Junior High School Mathematics Olympiad.

ABSTRAK

Olimpiade matematika tingkat SMP merupakan salah satu kegiatan rutian setiap tahunnya di Indonesia. Analisis butir instrumen tes Olimpiade matematika telah dilakukan oleh beberapa peneliti dengan menggunakan konsep teori tes klasik. Teori tersebut memiliki banyak kelemahan jika dibandingan konsep Item Response theory (IRT). Tujuan dari artikel ini yaitu mendeskripsikan analisis butir tes instrumen tes pra olimpiade matematika tingkat SMP di Pacitan Jawa Timur berbasis IRT. Analisis ini dilakukan dengan metode deskriptif kuantitatif. Analisis butir berbasisis IRT dilakukan dengan menggunakan software R dengan paket irtawsi. Hasil analisis dengan paket tersebut menunjukkan bahwa instrumen yang digunakan fit pada model 2PL. dari 15 butir yang digunakan tedapat dua butir yang tidak fit yaitu butir 4 dan 9. Selanjutnya dari 13 butir tes yang fit terdapat 3 butir yang tidak memenuhi kualitas daya beda yang digunakan. Tiga butir tersebut yaitu butir 1, 6 dan 13, sehingga tiga butir ini harus di drop karena dapat memberikan informasi yang bias jika digunakan untuk estimasi kemampuan olimpiade matematika siswa. Berdasarkan informasi tersebut dapat disimpulkan bahwa terdapat 10 butir tes yang memiliki kualitas yang memenuhi kualifikasi parameter kesulitan dan daya beda dengan standa error sebesar 0.25.

Kata kunci: IRT, parameter butir, Kualitas Butir, Olimpiade matematika SMP

Received	: June 16, 2023	/Revised	: September 3, 2023
/Accepted	: November 9, 2023	/ Published	: November 30, 2023

Introduction

Mathematics is one of the fields being contested in the OSN National Science Olympiad. Mathematics Olympiad is a means of improving the quality of education and is an event to find the seeds of outstanding students in the field of mathematics (Prawoto et al., 2019). Furthermore, the Ministry of Education and Culture emphasized that OSN is a strategic platform to equip students with the ability to think logically, systematically, analytically, critically and creatively and prepare students to master and create technology in the future (Kemendikbudristek, 2023). Those abilities are included in high-level thinking skills. Some of these abilities are also explained in (Akhsani & Purwanto, 2015; Setiani et al., 2022). Therefore,

it can be said that OSN seeks potential achievers through systematic logical thinking in solving higher-order thinking skill cases.

In the context of coaching students before participating in the 2023 OSN, the Pacitan Regency Mathematics Subject Teacher Consultation (MGMP) held a Pre OSN which could be attended by all class VII and VIII students from all junior high schools within the Pacitan Regency Education Office. In order to obtain reliable data and information for interested parties, the test instrument used must meet the criteria for a good test.There are two ways that can be used to determine the quality of tests, namely qualitative and quantitative analysis (Kustriyono, 2004). A qualitative approach is carried out by examining the questions before the test kit is tested. What is emphasized is the assessment of the material, construction, and language aspects. In contrast, the quantitative approach is a method of studying questions based on empirical data obtained through the responses of test takers. The quantitative approach is carried out after the test is given to the test takers and item information will be obtained which includes the level of difficulty, discriminating power and the effectiveness of the detractor. A good item must have an adequate level of difficulty, good discriminatory power, and function as a decoy (Suwarto et al., 2019).

Quantitative analysis on Olympic-type questions has been done before. Some study had conducted an Olympic-type test item analysis using the concept of classical test theory (Classical Test Theory or CTT) (Dewi et al., 2019; Ugi & Ekawati, 2016). The weakness of CTT is that the results of calculating the parameters of the difficulty level of the questions, discriminating power, and the reliability coefficient depend on the characteristics of the test takers as well as being influenced by the existing questions or items(Cappelleri et al., 2014). In addition, CTT is not an item analysis method that is able to show the relationship between item parameters and latent ability parameters that will be measured from Olympiad participating students.

Item analysis methods that are able to make items have functionality in measuring latent abilities can be done with item response theory (Items Response Theory or IRT). The ability of this method can be seen directly from the mathematical model function of IRT which shows that there is a direct relationship between the item parameters and the students' ability parameters (Retnawati, 2014). The advantages of using IRT in item analysis compared to CTT are (1) IRT is able to show errors in every measurement made by each item and CTT is not able to do that. (2) in IRT the test item parameters and ability parameters have an invariance property, and in CTT they do not have this property. (3) IRT has many other advantages over CTT (Eleje et al., 2018). Based on the explanation above, this article aims to carry out an item analysis of the junior high school mathematics pre-Olympiad test instrument in Pacitan district using the IRT concept which has never been done before for Olympic questions in Pacitan in particular. The benefit of this study is to obtain item parameters that can be used to estimate students' mathematics olympiad abilities.

Research Methods

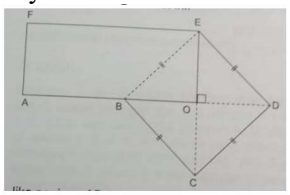
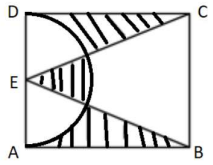
This study was conducted to describe the characteristics of the test items on the mathematics pre-Olympic questions in Pacitan district by using a quantitative approach. The pre-Olympic mathematics instrument used is the instrument used by the junior high school mathematics MGMP in Pacitan district. These instruments can be seen in table 1. According to data from MGMP, it was found that 55 junior high schools in Pacitan sent their candidates to participate

p-ISSN 2477-409X, e-ISSN: 2549-9084 and website: http://jurnal nasional.ump.ac.id/index.php/alphamath/

in this test. A total of 207 students participated, consisting of 169 (81.64%) female students and 38 (18.36%) male students.

Instrument item analysis was carried out using the IRT concept. Calibration calculations are carried out using the irtawsi package (Susanto et al., 2023). The steps used for the analysis were adopted from (Retnawati, 2014; Susanto et al., 2023; Susanto & Retnawati, 2023) namely (1) Determine the best model based on the Maximum likelihood method (2) Unidimensional assumption test, (3) Assumption test Local independence. (4) proof of item invariance, (5) Determining item Fit and (6) interpretation of the Information function.

Table 1. Mathematical Pre-Olympiad Test Instruments

No	Items
1	The remainder of the division is 217^{2022} by 10 is ….
2	The five numbers namely 1, 2, 3, 4 and 5 can all be arranged without repetition into 120 different numbers. If the numbers are sorted from smallest to largest, then the number that ranks 75th is...
3	If x and y are real numbers that satisfy $x^2 + y^2 = 1$, then the largest value of the multiplication of x and y is...
4	The average score of 25 students is 40. If the difference between the average score of the lowest 5 students and the average score of the other 20 students is 25, then the average value of the lowest 5 students is ....
5	Two bottles of the same size are full of saline solution. The ratio of salt and water content in the first bottle is 4:9 and in the second bottle is 3:7. If the contents of the two bottles are mixed, the ratio of the salt and water content of the resulting mixture is...
6	Consider $\frac{p}{13} + \frac{p}{3} = \frac{34}{39}$ If p and q are natural numbers, then the value of $p + q$ is ….
7	If it is known that the length of the edge of the cube ABCD.EFGH is one unit, then the distance from point E to the plane AFH is … units
8	A teacher wants to choose 4 out of 5 students and 4 students for the debate team. If at least 1 student and 1 female student must join a team, then the number of ways to choose is....
9	Pay attention to the following picture!  If the length of AB = 11 cm, BC = 15 cm and EF = 20 cm, then the area of ABCDEF is....
10	Is known $A = \{0,1,2,3,4\}$; a, b, and$c$ are three different numbers with A, as well as the value $(a^b)^c = n$. The maximum value of n is....
11	Two circles L1 and L2 have radii of 12 cm and 5 cm respectively. Point *P1* on *L1* and point *P2* on *L2*. First *L1* and *L2* intersect externally at *P1* and *P2*. Then *L2* is rolled along *L1*, so that it remains in external contact. Point *P2* first meets *P1* again when *L2* has been rolled ... times.
12	Pay attention to the pictures  On the quadrilateral ABCD, a semicircle is made on the AD side with center E and an equilateral triangle BEC. If BC = 20 cm, then the area of the shaded region is....
13	Calculate the value of $\frac{1}{1\times3} + \frac{1}{3\times5} + \frac{1}{5\times7} + \cdots + \frac{1}{2021\times2023}$
14	If $f(x+1) = 2f(x)$ and $f(1) = 5$, then the value of $f(7)$ is ….
15	The area of a rectangle is 3^{20} cm^2 and length 3^{22} cm, then the width of the rectangle is...

The parameter item difficulty level is set according to the default from the irtawsi package, which ranges from -4 to 4 (Susanto et al., 2023). Furthermore, the different power parameter adjusts the results of the IRT analysis and should not be less than 0.5. If there are items that have difficulty parameters outside the range of -4 to 4 or have different power parameters less than 0.35, then these items must be dropped. These items must be dropped, because it can cause bias if the parameters of these items are used to estimate ability scores.

Result and Discussions

Determine Model Fit

The results of calculations using the irtawsi package obtained the AIC, SABIC, HQ, and BIC values as in table 1. Determination of the best IRT model in this article is determined using the AIC coefficient.

Table 2. Selecting a Fit Model

Model	AIC	SABIC	HQ	BIC	logLik
quick	3844,027	3846,655	3865,590	3897,350	-1906,013
2PL	3762,869	3767,798	3803,301	3862,851	-1851,435
3pl	3765,702	3773,094	3826,349	3915,674	-1837,851

Based on table 2, the smallest AIC value is owned by the 2PL model, so that the instrument fits the 2PL model. This model is used as a basis for testing IRT assumptions.

Proof of Unidimensional Assumptions

The calculation results*irtawsi* shows that the MSA value = 0.725 which indicates that factor analysis can be used for unidimensional testing. The results of calculating the variance factor that can be explained are 47.6% > 20%. These results explain that the instrument measures only one lanten trait(Retnawati, 2014).

Proof of Local Independence Assumption

Testing the assumption of local independence is ignored according to the opinion of Retnawati, (2014), that is, if the assumption of unidimensionality is met, then the assumption of local independence is automatically fulfilled.

Invariansi Parameter

The results of calculating the parameter invariance proof can be seen in Table 5.

Table 3. Parameter Coefficient Correlation

Parameter	Correlation	p_Value	Interpretation
Discriminant	0,906	0	The assumption of invariance of the differential power parameters is fulfilled
Difficulty	0,824	0	The difficulty parameter invariance assumption is met
Ability	0,49	0	The ability parameter invariance assumption is met

Table 3 explains that there is no violation of the parameter invariance assumption. Both in item parameters and abilities. Based on the proof of these IRT assumptions, the 2PL model is feasible to use and proceed with item parameter estimation, determining item fit, and test information function.

p-ISSN 2477-409X, e-ISSN: 2549-9084 and website: http://jurnal nasional.ump.ac.id/index.php/alphamath/

Item Characteristics

The results of testing the characteristics of the items are presented in table 4 below.

Table 4. Fit Items

Item	Discriminant	Difficulty	ChiSQR	Item Fit
1	0.1	10.66	0.266	fit
2	2.4	0.44	0.279	fit
3	0.6	0.82	0.571	fit
4	0.8	-0.97	0.030	Not fit
5	1.5	0.64	0.614	fit
6	0.1	3.69	0.762	fit
7	1.2	1.10	0.260	fit
8	0.5	2.36	0.332	fit
9	0.7	0.95	0.030	Not fit
10	0.5	0.49	0.148	fit
11	0.9	0.51	0.400	fit
12	1.8	0.22	0.628	fit
13	-0.2	-4.25	0.202	fit
14	2.4	0.26	0.097	fit
15	0.9	0.78	0.092	fit

Based on table 4 above, there are two items that are not fit so they must be dropped or not used, namely item 4 and item 9. Furthermore, based on the value limit of the level of difficulty and differential power, item 1, 6 and 13 are also not used. The total items that can be used are 10 test items. The results of the Pre-Olympic instrument analysis using the IRT above explain that the instrument fits the 2PL model. The characteristics of this model are characterized by the presence of 2 parameters, namely the parameter of differential power and difficulty. The discriminating power parameter will provide an overview of the extent to which the items can differentiate students' abilities and the difficulty parameter will provide an overview of the difficulty level of the items.

Based on the difficulty level criteria in table 2. Then the difficulty level parameters in table 6 can be categorized into three, namely difficult, medium and easy. Items 1,3,6,7,8 and 15 are difficult items. Item 1 is an item that fits the model but the value of the difficulty level exceeds 4. Item 1 must also be dropped because it will cause bias if used to predict student abilities. Furthermore, items 2, 5, 10, 11, 12, and 14 fall into the medium category. In the easy category there is one item that is included in the easy category, namely item 13. The ability of the items to distinguish students who can work on a test item can be determined by categorizing in table 3. The first item has a high level of discriminatory power, namely 2,5,12, and 14. These items will be very selective in differentiating students' abilities. Second, items 7, 11, and 15 have the ability to distinguish students in the medium category. Third, points 3, 8, and 10 are able to distinguish students' abilities in the low category (Wulandari et al., 2020). Furthermore, items 1, 6, and 13 have very low discriminating power. These last three items should be dropped because they provide low information regarding student abilities. more extreme, item 13 with a negative discriminating power illustrates that low ability students have a greater chance of being able to do item 13 compared to high ability students. The higher the discriminating power, the more accurate item information will be in discriminating abilities (Hays et al., 2000).

Referring to the results of the difficulty level categorization and the different power of the Olympic test items, it can be concluded that out of the 15 items used only 10 items had good quality to measure students' mathematical olympiad abilities. These items are items 2, 3, 5, 7, 8, 10, 11, 12, 14, and 15. Furthermore, items 1, 6, and 13 are not used because they can provide biased predictions of students' Olympic ability. Apart from these three items, items 4 and 9 are two items that are not fit. These two items are not suitable for use in predicting students' ability to answer the given Olympic questions. Based on 10 items that can be used to predict students' Olympic abilities, the information function is obtained in Figure 1.

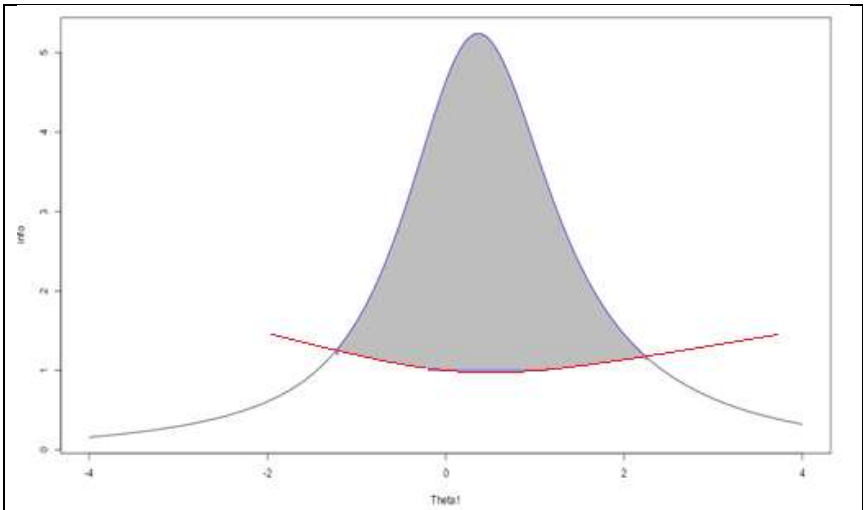


Figure 1. Information Function and Error Standar curve

The shaded area in figure 1 shows the functioning of the 10 test items of the Mathematics Olympiad Pra instrument in Pacitan. Based on the intersection of the information function and the standard error, the instrument is suitable for measuring math olympiad abilities between 1.25 to 2.27.

Conclusion

Refer to the results and discussionit can be concluded that of the 15 test items, there are 13 fit items. Of the 13 items there were 3 items that did not meet the discriminant limits used, so there were only 10 items that had good quality.

Acknowledgement

The author would like to express sincere gratitude to the Mathematics Teachers' Working Group (MGMP) of Pacitan District Junior High Schools for providing the data source that has been instrumental in the completion of this research.

Bibliography

Akhsani, L., & Purwanto, J. (2015). Upgrading Higher-Order Thinking of UMP Mathematics Education Student Through Project Based Learning Model in Advanced Calculus Course 1. *AlphaMath : Journal of Mathematics Education*, *1*(1). https://doi.org/10.30595/alphamath.v1i1.207

Cappelleri, J. C., Jason Lundy, J., & Hays, R. D. (2014). Overview of Classical Test Theory and Item Response Theory for the Quantitative Assessment of Items in Developing

p-ISSN 2477-409X, e-ISSN: 2549-9084 and website: http://jurnal nasional.ump.ac.id/index.php/alphamath/

Patient-Reported Outcomes Measures. *Clinical Therapeutics*, *36*(5), 648–662. https://doi.org/10.1016/j.clinthera.2014.04.006

Dewi, S. S., Hariastuti, R. M., & Utami, A. U. (2019). Analisis Tingkat Kesukaran Dan Daya Pembeda Soal Olimpiade Matematika (Omi) Tingkat Smp Tahun 2018. *Transformasi : Jurnal Pendidikan Matematika Dan Matematika*, *3*(1), 15–26. https://doi.org/10.36526/tr.v3i1.388

Eleje, L. I., Onah, F. E., & Abanobi, C. C. (2018). Comparative study of Classical Test Theory and Item Response Theory using diagnostic quantitative economics skill test item analysis results. *European Journal of Educational & Social Sciences*, *3*(1).

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item Response Theory and health outcomes measurement in the 21st century. *Medical Care*, *38*(9 SUPPL. 2). https://doi.org/10.1097/00005650-200009002-00007

Kemendikbudristek. (2023). *Pedoman Olimpiade Sains Nasional (OSN) Jenjang SMP/MTs Tahun 2023*.

Kustriyono, K. (2004). Penyusunan Perangkat Soal Ujian Akhir Mata Pelajaran Sains-Biologi SMP dalam Rangka Pengembangan Bank Soal. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *6*(2). https://doi.org/10.21831/pep.v6i2.2048

Prawoto, B. P., Sulaiman, R., Savitri, D., & Fardah, D. K. (2019). Pelatihan Pendamping Olimpiade Matematika Smp Kabupaten Tulungagung. *Jurnal ABDI*, *5*(1), 21. https://doi.org/10.26740/ja.v5n1.p21-24

Retnawati, H. (2014). *Teori Respons Butir Dan Penerapannya: Untuk Peneliti, Praktisi Pengukuran Dan Pengujian, Mahasiswa Pascasarjana.* Yogyakarta: Nuha Medika.

Setiani, N. W., Asikin, M., & Dewi, N. R. (2022). Numerical Literacy Skills of Vocational High School Students in Solving HOTS Problems. *AlphaMath : Journal of Mathematics Education*, *8*(2), 121-130. https://doi.org/10.30595/alphamath.v8i2.14161

Susanto, H. P., & Retnawati, H. (2023). Kalibrasi Instrumen Literasi Matematika Siswa Menggunakan IRT dan Aplikasinya untuk Estimasi Skor. *Edumatica: Jurnal Pendidikan Matematika*, *13*(1), 23–36. https://doi.org/10.22437/edumatica.v13i01.23135

Susanto, H. P., Retnawati, H., Abadi, A. M., Haryanto, H., & Ali, R. M. (2023). *irtawsi: Items Response Theory Analysis with Steps and Interpretation* (R package version 0.3.4). CRAN R Pgroam. https://cran.r-project.org/package=irtawsi

Suwarto, Widoyoko2, E. P., & Setiawan, B. (2019). The Effects of Sample Size and Logistic Models on Item Parameter Estimation. *Proceedings of the 2nd International Conference on Education*, 323–330.

Ugi, L. E., & Ekawati, D. (2016). Kualitas Tes Pra Olimpiade Bidang Studi Matematika tingkat SMPdi Kota Baubau. *Prosiding Seminar Nasional*, 118–127. https://journal.uncp.ac.id/index.php/proceding/article/view/379

Wulandari, F., Hadi, S., & Haryanto, H. (2020). Computer-based Adaptive Test Development Using Fuzzy Item Response Theory to Estimate Student Ability. *Computer Science and Information Technology*, *8*(3). https://doi.org/10.13189/csit.2020.080302