5th ICRIEMS Proceedings
Published by Faculty Of Mathematics And Natural Sciences
Yogyakarta State University, ISBN 978-602-74529-3-0

# Analysis Characteristic of Diagnostic Instrument to Measure Error of Mathematics Problem Solving based on Politomus

Muhamad Arfan Septiawan[1,a)], Heri Retnawati[2,b)]

[1] *Program Graduate School of Yogyakarta State University*
*Kolombo Street No. 1 Karang Malang, Depok, Sleman, Yogyakarta, Indonesia*
[2]*Faculty of MIPA of Yogyakarta State University*
*Kolombo Street No. 1 Karang Malang, Depok, Sleman, Yogyakarta, Indonesia*

a) smarfan020@gmail.com
b) retnawati_heriunyI@gmail.com

**Abstract**. Instrument is a tool used to measure a measuring object or data collection from a variable. The better instrument will produce the better data. This study aims to determine the characteristics of instrument test diagnostic of mathematics problem solving based on politomus. This study use quantitative approach. Population in this study is all of accredited state senior high school in Lampung Utara Regency. Schools are divided into three categories A, B and C accreditation. The sample of the study amount 203 students taken using Stratified Random Sampling technique that represent each level. Instrument of test totaling 20 items in the scored politomus Partial Credit Model (PCM). The result of this study is information of validity, reliability, difficulty level, suitability of item (Item Fit), gender bias, domicile bias and combination of two demographics. The results showed that the instrument test diagnostic has fulfilled the validity of the content qualitatively with expert judgment, quantitatively acquired the Aiken index was 0.844 (High category). Reliability index is obtained with the WINSTEPS program. Person reliability index was 0.63 (weak category) item reliability index was 0.94 (excellent), reliability index of interaction between person and item as a whole with Cronbach Alpha value was 0.68 (enough). These has met suitability of item (item fit), and all items fit. The difficulty level used logit value, the highest logit value shows the highest difficulty level was S18 with logit value + 0.37 and the easiest item was S1 with logit value -0.42. All items in the diagnostic test are in the moderate category difficulty level at the range of $-0{,}37\ to + 0{,}42$. Thebias gender, domicile and combination of two demographics are known from the value of DIF > 0.05 by looking at the probability value, and got 2 gender bias questions were $S6 = 0.0037$, $S7 = 0.0073$, got 1 demographics bias questions were $S6 = 0.0197$.

## INTRODUCTION

In the world of education the ability of teachers in making a good test instrument becomes an obstacle to obtain information in measuring the ability of students as a whole in the mastery of mathematics [1]. Most teachers prepare the test instruments based only on indicators or outcomes with MGMP members of the mathematics course without analyzing the characteristics of the item to obtain good instruments and measurement results. Analysis question is part of teacher administration that must be implemented. The task of teacher administration is not only to make the question and then to assess the student's work and move to the list of values, but it is also important to carry out a characteristic analysis of the question.

In general, the analysis used usually includes three information about the characteristics of test items, namely the level of difficulty, differentiation, distractors for data dichotomy [2]. In addition the validity and reliability of the tests can be used to strengthen the information obtained [3],[4],[5]. The validity of a test instrument is the ability of a test to measure what is actually measured [6], [5], [7],[2]. While reliability is the stability score obtained by the same person when retested and reliability is a very important thing is owned by a test or instrument [8].

Most teachers and researchers are often still doing the analysis using the classical way [21]. Few of the teachers and researchers analyzed using the item response theory model (IRT) [20][18]. Some researchers have

conducted research related to the characteristics of test instruments. [4] Analyzed the validity and reliability of basic mathematics test instruments related to the introduction of core physics with the result of validity was high criteria 0.77. Instrument reliability was 0.87 is very high category. [9] Conducted reliability analysis and construct selfconcept scale validity for Indonesian students with composite reliability coefficient was 0.98.

However, in assessing the characteristics of a good instrument is not sufficient only on the basis of general analysis. Therefore, there are several characteristics that need to be raised, among which are the characteristics of person reliability, item reliability characteristics, reliability index of interaction between person and item, suitability item (item fit) and bias (DIF) [10]. Analysis of reliability characteristics in person and item is still very rarely raised by most researchers, suitability of item (item fit) and gender bias, domicile bias and combination of demography becomes something important to come up with.

In this paper we will discuss the characteristics of the Partial Credit Models (PCM) -based question model for qualitative validity through expert and quantitative opinions based on the Aiken formula [11]. Looking for individual reliability, item reliability, and reliability of interaction between personnel and items together as well as difficulty level, suitability of item (item fit), and gender bias with Rasch modeling using the help of WINSTEPS program.

## METHODOLOGY

This research is a quantitative descriptive research to provide information about diagnostic test instrument characteristics of problem solving based on politomus. Sampling was done by stratified random sampling technique to determine three schools based on accreditation. In determining the number of samples was done by purposive sampling by choosing one class of natural science and social science in each School. This research was conducted on February-March 2018 at State Senior High School class X natural science and social science in North Lampung Regency, Lampung Province.

Population in this research as many as 9 Senior High School located in Lampung Utara with accreditation A, B and C. With stratified random sampling technique choosen 3 State Senior High School representing each accreditation that will be made as research sample. Determination of the number of samples by purposive sampling by choosing one natural science and social sience class from each school. Number of sample is 203 students.

Validity has a big role in the characteristics of the question[12] Validity was done qualitatively and quantitatively, for qualitatively using the expert opinion while for quantitative using Aiken formula [11], [5]. The data was collected by giving 20 items in the form of multiple choice based on partial credit Model (PCM) model scoring to 203 students.Analysis of item test was performed by modeling Rasch [13] through the help of the WINSTAP software to know the characteristics of the test items including difficulty level, person reliability, reliability of combined items of reliability, suitability of item (fit items), gender bias, domicile bias, and combination of two demographicsInformation on difficulty levels is seen by logit values, high logit values indicate the highest difficulty level [14].

In Rasch modeling to see the person reliability, the reliability of the combined items of reliability can be seen through statistical summary and test information functions. Criteria value of person reliability and item reliability is if the reliability value <0.67 = weak, 0.67-0.80 = enough, 0.8-0.90 = good, 0.91-0.94 = very good,>0.94 = excellent. Cronbach alpha value to measure reliability, that is interaction between person and item question overall with criteria if value of reliability <0,5 = bad, 0,5-0,6 = ugly, 0,6-0,7 = enough, 0.7-0,8 = good and> 0.8 = very good [14].

In Rasch modeling beside difficulty level item and statistical summary is to see the suitability of item with model (Item Fit). Outfit means-square value, z-standard outfit and point measure corelation are the criteria used to assess the suitability of the items level[15], [16], [14]. According to Boone et al [15] the criteria used to examine the outfit mean square (MNSQ) received if 0.5 <MNSQ <1.5. The accepted Z-standard Outflit (ZSTD) value -2.0 <(ZSTD) <+2,0. Point Measure Corelation Value (Pt Mean Corr) 0.4 <Pt Measure Corr <0.85.

According to Sumintono & Widhiarso [14]from the three criteria above if the value of Outfit MNSQ and Point measure correlation does not eligible the requirements, but for the criteria of Outfit ZSTD value is still within the allowed limit then the item does not need to be changed or replaced. Beside through the numbers, the mismatch of item can be displayed through the ICC's expected score chart.In Rasch modeling to detect bias was called DIF (differential item functioning) detection. A question item was said to contain bias if found that its probability value is below 5% or 0.05 [14]. Beside to the numbers, information on the question item having a bias (DIF) is also raised with the Person DIF plot graph.

# RESULT AND DISCUSSION

**a. Validity of The Diagnostic TestItems**

The result of validity of diagnostic test instrument of problem solving based on scoring politomus model of partial credit model (PCM) in class X State Senior high school in North Lampung Regency qualitative study by three experts [3][18], one lecturer of postgraduate mathematics education program, one lecturer Faculty of Mathematics and Natural Sciences Mathematics Education and Mathematics Teacher of State senior high school 2 Kotabumi Lampung Utara. Quantitatively using Aiken formula obtained Aiken index of 0.848 in the high category.

**b.Reliability Diagnostic Test**

In Rasch modeling to see the person reliability, item reliability, combined reliability can be seen through statistical summary and test information function. Criteria value of person reliability and item reliability is if the reliability value <0.67 = weak, 0.67-0.80 = enough, 0.8-0.90 = good, 0.91-0.94 = very good,> 0.94 = excellent. Cronbach alpha value to measure reliability, that was interaction between person and itemsoverall with criteria if value of reliability <0,5 = bad, 0,5-0,6 = ugly, 0,6-0,7 = enough, 0.7-0,8 = good, and> 0.8 = very good [14].

On the reliability aspect can be seen from the output of statistical summary and function of measurement information in WINSTAP program. Based on the output of statistical summary of the WINSTAP program obtained Person reliability value was 0.63 (weak category) wih the value of separation 1.30, while the reliability index item of 0.94 (very good) with the value of separation 3.94. In the interaction aspect between person and item, the cronbach alpha index is 6.8 (enough).

**c. Problem difficulty level**

Information of difficulty levels was seen based on logit values [10], high logit values indicate the highest difficulty level [14] The level of difficulty (b) in the item response theory was said to be good if it lies at interval [-2.2] [2]. Based on the logit criteria, the level of difficulty can be grouped into 3 categories:1). Hard; with an index of difficulty> 2.002). Medium; with an index of difficulty -2.00 to 2.00, 3). Easy; with an index of difficulty <-2.00.

To know the difficulty level of the question can be seen outfit measure items from WINTSAP program[14].

The items in the table are sorted by the logit value of the largest (ie the 18th) to the smallest logit (ie the 1st). For question number S18 is the highest logit value that is +0.37 which explains the number S18 is the hardest problem. While the question number S1 is a problem with the lowest difficulty level with a logit value of -0.42. Based on the above results can be concluded has three levels that is hard, moderate, easy. will be presented in the following table

**Table 4**. difficulty level of item based on Rasch modeling

| Category | Mathematics diagnostik item | |
|---|---|---|
| | Item number | Count |
| easy ($b < -2$) | S11, S10, S8, S6, S9, S1 | 6 |
| moderate ($-2 \leq b \leq 2$) | S2, S13, S20, S3, S17, S14, S12, S5, S7, S19, S4, S15, S16, S18 | 14 |
| Hard ($b > 2$) | - | 0 |

Based on a summary of outfit measure All items contained in the diagnostic test instances are at moderate category difficulty levels in the range of -0.42 to 0.37.

**d. Suitability Level of Problem Item**

In Rasch modeling besideitem difficulty level and statistical summary is to see the suitability of the item with the model (Item Fit). Outfit means-square value, z-standard outfit and point measure corelation are the criteria used to assess the suitability of the items [15], [16], [14]. According to Boone et al [15] the criteria used to examine the suitability of the items that are not suitable can be seen value of Outfit mean square (MNSQ) accepted if 0.5 <MNSQ <1.5. The accepted Z-standard Outflit (ZSTD) value was -2.0 <(ZSTD) <+2,0. Point Measure Corelation Value (Pt Mean Corr) was 0.4 <Pt Measure Corr <0.85.

In the table above shows that the top item is S6 has a tendency not fit. When viewed from all three criteria, item S6 only fails to qualify at Mnsq outfit with value 1.17 and point measure corelation with value 0,40, but for criterion outfit Zstd value still allowed. According to Sumintono & Widhiarso [14]from the three criteria above if the value of Outfit MNSQ and Point measure correlation fail qualify the requirements, but for the criteria of Outfit ZSTD value is still within the allowed limit then the item does not need to be changed or replaced.

Value of ZNSQ item S6 (1,4), S14 (1,7), S13 (1,4), S9 (0,4), S15 (1,5), S4 (1,3), S11 (0,9 ), S7 (0.8), S19 (0,5), S12 (0.4), S1 (0.3), S5 (-0.2), S2 (-0.2), S16 (-0 4, S20 (-0.7), S17 (-0.9), S8 (-1.0), S10 (-1.4) S18 (-2.0) and S3 (-1.9) . Thus all items can be used and fit the fit model, without needing to be replaced or changed as it is within limits.

**e. Detection of bias items**

In Rasch modeling to detect this bias is called detection of DIF [14], [2], [10]. Consider also the functioning of the differential point so that it does not favor one of the groups [2]. A question item is said to contain bias if it is found that the probability value of the item is below 5% or 0.05 [14], [10]. Beside to the numbers, information of a biasitem (DIF) is also raised with a Person DIF plot graph.

Based on the probability value of 20 items we get two items about the gender bias that has a value below 0.05 that is about the number S6 with probability value (0.0037), and the question number S7 with probability value 0.0076).

In addition to using probability values, gender bias can also be viewed through the following person plot of the DIF plot.

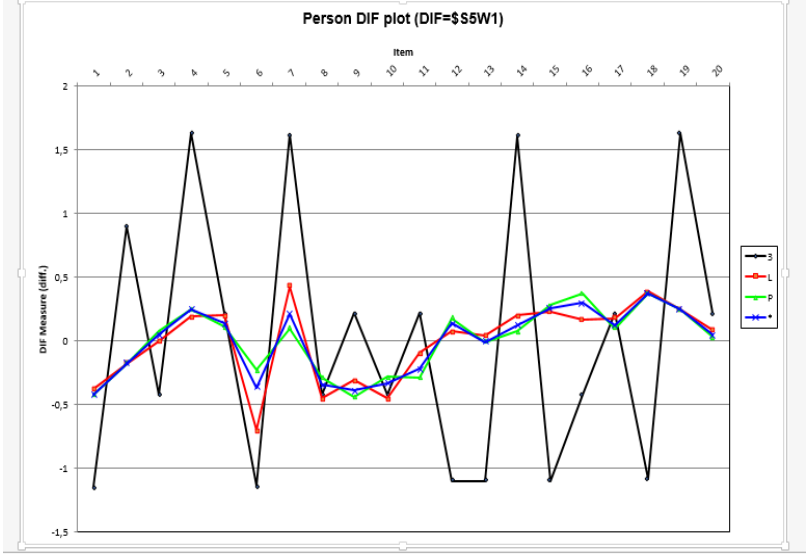


**Figure 1.**Graph Detection Of Bias Items

On the graph the curve seen approaching the upper border of item 7 indicates the highest difficulty level. While the curve is below as item number 6, it shows an easy point item. About the item DIF, there are two numbers that contain the DIF number 6 and 7. It is seen that item 6 is more easily handled by male (below), compared to female students with far differences. While the number 7, a difficult problem can be done by male than female. For other items, the difference in the ability to work on items correctly does not vary much.

In addition to looking at gender bias, in this paper DIF will be show domiciled students who live in the city districts and in the village areas are shown by the probability value and graph of the person DIF plot.

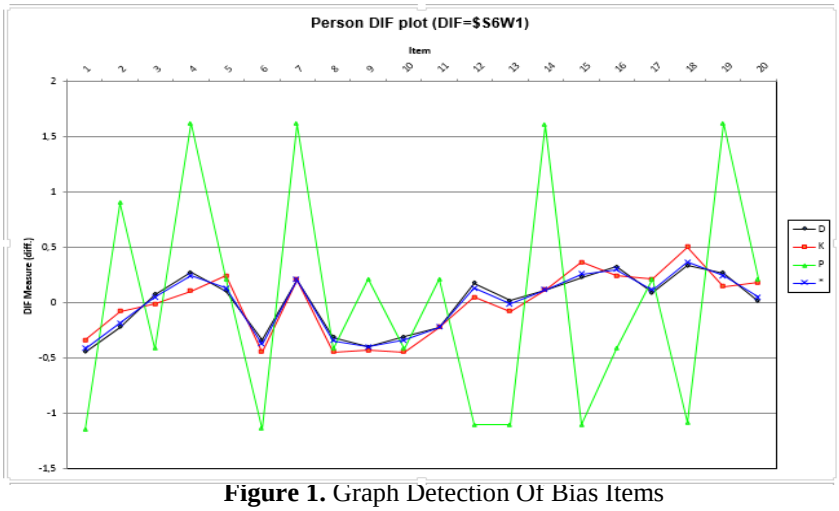


**Figure 1.** Graph Detection Of Bias Items

Based on the probability value of 20 items, all probability values> 0.05 so there is no bias in the student's domicile. Based on the graphic, the differences in the ability to work on items are not different between students living in the city and in the village.

The DIF analysis in this paper will combine two demographic data. That is the type of sex (L and P) and the domicile combination (D and K). We can look at which combinations of items that contain DIF in a particular demographic group. The DIF information will be presented via the probability value and the DIF plot person graph as follows.

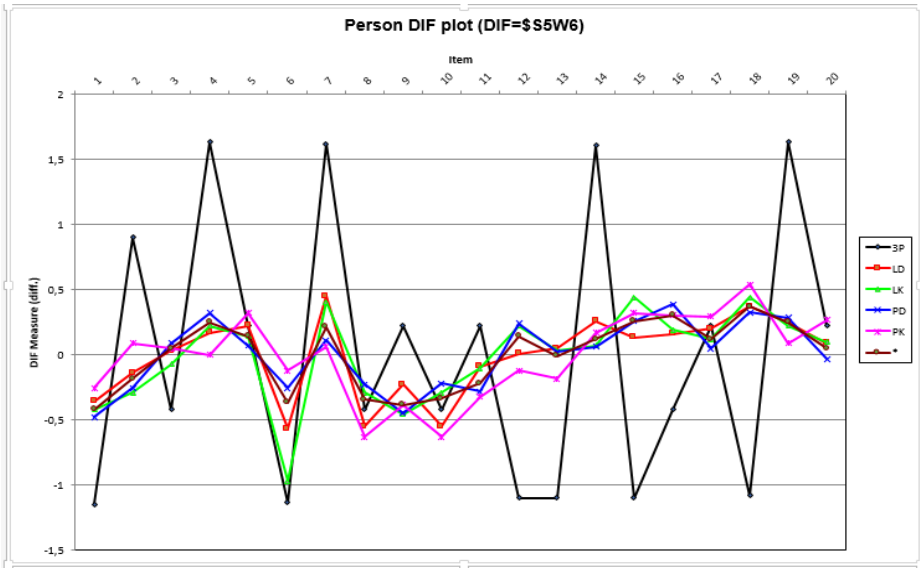


**Figure 1.** Graph Detection Of two demographi

Based on the probability value of 20 items, obtained the number S6 bias with probability value 0,0197 <0,05. While based on the graph it is seen that item 6 is easier to be done by village male (LD) than by urban male students (LK), village female (PD), city female (PK).

The analysis of the DIF curve on the graph can provide information for a particular group. For the LD, (male gender domiciled in the village) it can be explained that item 7, 9, 14 are includng a difficult question and can be answered correctly than any other group. while the numbers 15 and 16 are a matter of items that are easy and can be done correctly than any other group.For the category of LK (male gender domiciled in the city), it can be explained that item 15 is a difficult question and can be answered correctly compared to other groups. while the numbers 6 and 9 are a matter of items that are easy and can be done correctly compared to other groups.For the category of PD (female gender domiciled in the village), it can be explained that item 3, 4, 8, 10, 12, 16, 19 are difficult questions and can be answered correctly compared to other groups. While the numbers 1, 5, 14, 17 and 20 are questions with items that are easy and can be done correctly than other groups.For the category of PK (female gender domiciled in the city), it can be explained that item 1, 2, 5, 6, 18, 20 are a difficult question and can be answered correctly compared to other groups. While the numbers 4, 7, 8, 10, 11, 12, 13, 19 are a matter of items that are easy and can be done correctly than any other group.

## CONCLUSION

Characteristic analysis of diagnostic test of problem solving based on Polytomus scoring type PCM, able to provide accurate and detailed information based on Rasch modeling with the help of WINSTAB program. In this paper, we have presented information about the characteristics of validity, person reliability, item realibility, or combined realibility, difficulty level,suitability item (item fit), gender bias, domicile bias or combination of two demographic data. Analysis using Rasch modeling with the help of WINSTAB program is very well used by researchers and teachers in analyzing the characteristics of an instrument.

## ACKNOWLEDGMENTS

The author thank to the Faculty ofMathematics and Science, Program Graduate School of Yogyakarta State University, and all of contributors for this work.

## REFERENCES

1. Retnawati, H., et al, (2018) Teachers' Difficulties and Strategies in Physics Teaching and Learning That Applying mathematics. *Journal of Baltic Science Education*, (17), (1) 2018.
2. Winarno. S & Kartowagiran. B (2013). Differential item functions in the grade promotion test of mathematics for grade VIII of junior high schools in sleman regency. *Journal of Evaluation Education*. No 2, vol 1, 2013.
3. Retnawati. H (2016). *Validitas Reliabilitas & Karakteristik Butir*.Yogyakarta: Parama Publishing.
4. Kereh et al.,2015. Validitas dan reliabilitas instrumen tes matematika dasar yang berkaitan dengan pendahuluan Fisika Inti. *Jurnal Inovasi dan Pembelajaran Fisika,(*2), (1), 2015.
5. Azwar, S. (2012). *Reliabilitas dan Validitas* (edisi 4). Yogyakarta: Pustaka Pelajar
6. Allen M.J. & Yen, W.M (1979). Introduction to measurement theory. Montery, CA: Brook/Cobe Publishing Company.
7. Kerlinger, F.N. (1986). *Asas-asas penelitianbehavioral.* (Terjemahan L.R. Simatupang). Yogyakarta: mada University. Pers
8. Kartowgiran, Badrun. (2005). *Perbandingan berbagai metode unuk mendeteksi bias butir*. Disertasi Doktor, tidak diterbitkan, Universiyas Gadjah Mada, Yogyakarta.
9. Widodo P.B (2010). Reliabilitas dan validitas konstruk skala konsep diri untuk mahasiswa indonesia . *Jurnal psikologi Universitas Ponorogo*(3)(1), 2006.
10. William J. Boone, jhon R. Staver, Melissa S. Yale, 2014. *Rasch Analysis in the Human Sciences.* Springer dordrecht Heidelberg New York London
11. Aiken, L.R. (1987). *Assesment of Intelectual Functioning*. Boston. Allyn & Bacon, Inc.
12. John A. Jhonson, 2004. *Mulivariance behavioral reasearch* ,39 (2), 273-302.
13. Benjamin D. Wright, 1978. The Rasch for test construction and person measurement. University of Chicago. Fifth Annual Conference and Exhibition.
14. Sumintono. B & Widhiarso. W, (2015). *Aplikasi Permodelan RASCH pada Aseeseement Pendidikan.* Bandung: trim Komunikata.
15. Boone, W., and Yale, M.S (2014). *Rasch Analysis in the Human Sciences*. Dordrecht: Springer
16. Bond, T.G., & Fox, C. M. (2015). *Applying the rasch model: fundamental measurement in the human sciences.* (2nd Ed.). Mahwah:Lawrence Erlbaum Associates, publishers.
17. Embretson, S.E (2007). Construct validity : A Universal Validity system of just another tes evaluation prosedure. Educational Researcher, 36 (8), 449-455.
18. Noverma.N., (2016). Analisis Kesulitan SELF-EFFICACY Siswa dalam Pemecahan Masalah Matematika Berbentuk Soal Cerita. *Jurnal Riset Pendidikan Matematika*. (3), (1), 2016.
19. Kumalasari. A & Sugiman., (2015). Analisis Kesulitan Belajar Matemtika pada Mata KuliahKapita Selekta Matematika Sekolah Menengah. *Jurnal Riset Pendidikan Matematika*(2). (1), 2015.
20. Retnawati, H., Hadi, S., & Nugraha, A.C. (2016). Vocational high school teachers' difficulties in implementing the assessment in Curriculum 2013 in Yogyakarta Province of Indonesia. *International Journal of Instruction,* 9(1), 33 – 48.
21. Retnawati.H, et al., (2017). Way are Mathematics National Examination Items Difficult and What Is Teachers' Strategy to Overcome It?. *International Journal Instruction.*(10). (3), 2017.