International Journal on Emerging Mathematics Education (IJEME)
Vol. 6, No. 1, March 2022, pp. 29-40
P-ISSN: 2549-4996, E-ISSN: 2548-5806, DOI: http://dx.doi.org/10.12928/ijeme.v6i1.21761

Rasch Model Study on Mathematics Examination Test Using Item Response Theory Approach

Vegi Intan Futri*, Raden Rosnawati, Abdul Rahim, Marlina
Universitas Negeri Yogyakarta, Jl. Colombo No. 1 Karangmalang, Depok, Sleman, DIY, Indonesia
*Corresponding e-mail: vegiintan.2019@student.uny.ac.id

Abstract

This study was conducted to analyze the test instrument used to measure the ability of students in the odd final exam in mathematics. Sampling using purposive sampling technique. These students consist of 398 students. The questions given are in the form of multiple-choice questions with a total of 40 items. The data analysis technique used quantitative descriptive analysis. Data analysis was carried out using the Item Response Theory (IRT) Rasch model approach with the help of QUEST software. The results of the analysis show, from 40 items there are 39 items fit with the Rasch model. Judging from the level of difficulty, items with very difficult categories are 0%, difficult categories are 8 items or 21%, moderate item categories are 23 items or 59%, easy categories are 8 items or 21%, and very easy categories are 0 %. The reliability of the estimated value of the item is 0.95 with a very good category so that it affects the items that fit the model. The higher reliability, the more items that fit the model. The reliability of the case estimate value is 0.00 with a weak category. This value indicates an inconsistency in the answers of the test takers, which means that the test takers are careless in answering the questions, thus affecting the reliability of the questions.

Keywords: Final Exam Test; Item Response Theory; Rasch Model.

How to Cite: Futri, V. I., Rosnawati, R., Rahim, A., & Marlina (2022). Rasch Model Study on Mathematics Examination Test Using Item Response Theory Approach. *International Journal on Emerging Mathematics Education, 6*(1), 29-40. http://dx.doi.org/10.12928/ijeme.v6i1.21761

INTRODUCTION

The learning process carried out in schools is still the responsibility of the teacher to continue to evaluate and assess learning (Anwar et al., 2019; Rahayuningsih & Jayanti, 2019). Appropriate assessment can help improve the learning process, so organizing learning tools or measuring tools is one of the most important things for a teacher (Naqiyah et al., 2020). Instruments as measuring tools can be in the form of tests or non-tests (Purwanto, 2009) one of them is a test instrument as a measuring tool for data collection to measure in the cognitive domain as a certain way that can be used or procedures that need to be taken in the context of measuring and assessing in the field of education (Kadir, 2015).

The instrument used as a good assessment will provide a higher information value than the measurement error (Marjiastuti & Wahyuni, 2014; Heri Retnawati et al., 2016). A high information value will provide an overview of the actual measurement results to provide high information, then the instrument to be used in measurement activities must be valid and reliable (Djemari, 2012; Hari Retnawati, 2014; Tri Wahyuningsih, 2015). The quality of an instrument that can be proven by looking at the accuracy of the instrument in measuring the validation that is intended to be the measurement goal (Alkharusi, 2015; Djemari, 2012; Kartianom & Ndayizeye, 2017; Pey Tee & Subramaniam, 2018; Hari Retnawati, 2014; Rindermann & Baumeister, 2015; Wu et al., 2016). The validity of an instrument can be proven in terms of content, constructs, and criteria (Djemari, 2012; Heri Retnawati et al., 2016; Wu et al.,

Received September 15, 2021; Revised November 22, 2021; Accepted November 24, 2021

2016). Instrument reliability related to reliability used in measurement activities in producing consistent information or results (Wu et al., 2016). The higher the value of the validity and reliability of an instrument, the more accurate the data obtained from a study will be (Hayati & Lailatussaadah, 2016). Measurement instruments also have characteristics that are described by the items of the instrument by conducting an empirical analysis (Heri Retnawati et al., 2016). The test instrument was used as a formative assessment. One way that is considered suitable for teaching and assessing competencies that supports the 21st century is using formative assessment strategies (Griffrin & McGaw, 2014; Shute & Becker, 2010; Wafubwa & Csíkos, 2020). That unlike summative assessment which is used as a measurement instrument, formative assessment is designed to be a useful assessment for teachers and students (Clark, 2012; Gipps et al., 2015). Formative assessment also serves as an opportunity for teacher professional development (OECD, 2005).

Empirical item analysis in classical test theory and item response theory (Fitriani et al., 2019; Siti Aminah, 2013). The level of difficulty and discrimination in classical test theory determines the quality of the items. The characteristics of the items produced by the classical test theory are inconsistent depending on the group of exami. This is then used by some experts as part of the weakness of the classical test theory approach known as group dependent (Pratama, 2020; Heri Retnawati et al., 2016). Weaknesses in the level of item difficulty and item discrimination depend on the group of exam. However, in reality, a person's ability to answer correctly from an item depends on the individual ability of the examinee itself, not based on the ability of the group of exam. That someone who learns and understands the subject being studied will be able to work on the questions well. This means that the chance of correctly answering the questions tested is higher than the participants who did not study.

Overcome the weakness of classical test theory in the measurement expert to develop a model that is not tied to the sample. This model is then known as the modern test theory. IRT has the assumption that the chances of test takers answering correctly for each item depend on the ability of test takers who have high abilities to have a greater chance of answering correctly than those with low abilities (Heri Retnawati & Hadi, 2014). The item response theory (IRT) approach is an alternative approach that can be used in analyzing a test and in the processes to obtain valid measurement instruments (Aricak et al., 2020). The IRT model is widely used by experts in developing tests, one of which is the Rasch model. Rasch model is very easy to do and apply with accurate analysis results (Susdelina et al., 2018) which examines the opportunity to answer correctly on the question by comparing the student's ability with the level of difficulty of the question (Sumintono, 2014) as the development of a data measurement Rasch model that can determine the relationship between the student's own level of ability (person ability) and the level of item difficulty by using the logarithmic function to be able to produce measurements with the same interval value. The selection of the Rasch model is because this model already has the principles of a measurement model, namely; able to provide a linear measure with the same interval, able to overcome the problem of missing data, can provide more precise estimates, can detect the imprecision of a model, and provide independent measurement instruments from the parameters studied (Abdullah et al., 2012; Sumintono, 2014).

Research conducted by (Santoso et al., 2019) that the items analyzed are able to provide accurate information using the Item Response Theory approach. Research conducted by (Kartianom & Mardapi, 2017) which utilizes the National Examination data which is analyzed with the item response theory approach to find out the strengths and weaknesses of students whose information can be used by the teacher as material for learning improvement and research conducted by the teacher. (Imaroh et al., 2017) Regarding item analysis using the Rasch model, it can provide information about the quality of the items in the final test of odd semester mathematics for grade VII Junior High School (SMP). Research conducted by (Susanto et al., 2015) to analyze the validity, reliability, level of difficulty and differentiating power on the odd semester final exam items for mathematics subjects. Research conducted by (Safihin, 2019) produce objective tests to measure student learning outcomes, the characteristics of items can be analyzed using the Rasch Model approach.

Based on the research that has been done, the researchers want to know the quality of the test instruments used to measure students' abilities in the final exams of the odd semester mathematics class VIII SMP with the Rasch model approach. The quality measured for several indicators includes items that fit the Rasch model, the level of difficulty of the items, and the reliability of the items designed by the test instrument which were then determined which items were fit and unfit with the Rasch model and determined the Cronbach alpha value to determine the items' reliability.

RESEARCH METHOD

This research is a quantitative descriptive study with the aim of focusing on the analysis of the odd semester final exam test instrument with the coverage of material including: Number Patterns, Cartesian Coordinates, Relations and Functions, Straight Line Equations, and Two Variable Linear Equations System. The analysis of the odd semester final exam test instrument uses the Rasch model approach. Sampling using purposive sampling technique. The subjects of this study were students of Junior High School (SMP) class VIII in Yogyakarta as many as 398 students. There are 40 multiple choice questions on the final semester exam test instrument aimed at students. Quantitative data analysis was carried out using the Rasch IRT approach with the help of the QUEST program.

RESULTS AND DISCUSSION

The final semester exam test instrument has 40 items with four choices. Analysis of respondents answer patterns were analyzed using the Rasch model through the QUEST software. The quality of the questions in the Rasch model can be known by estimating parameters such as validity, reliability, discriminating power, level of difficulty, and item fit with the Rasch model.

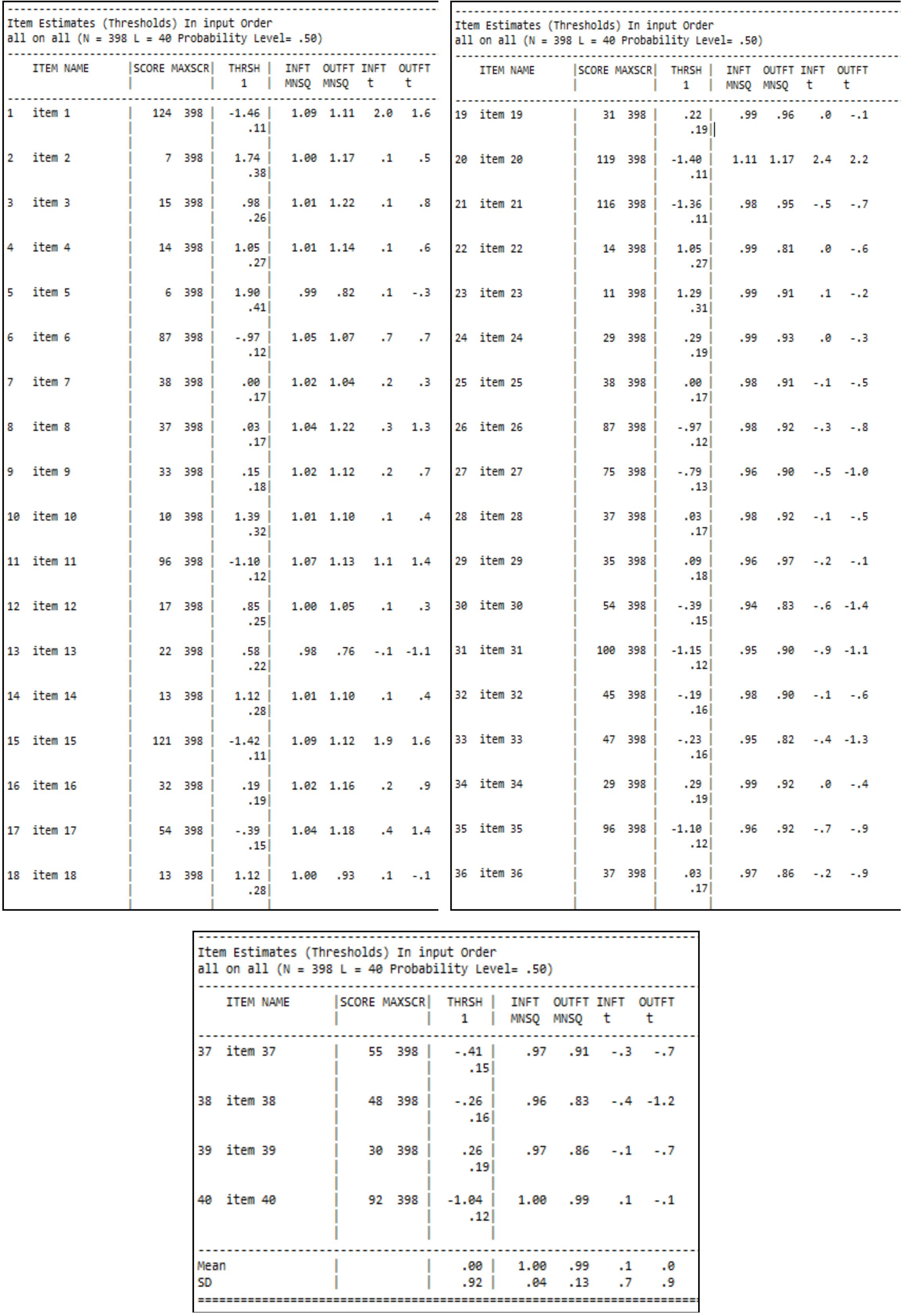
Estimated Item Validity

Test the validity using the QUEST program as disclosed Setyawarno, (2017) can be compared through the criteria in Table 1.

Table 1. Criteria for INFIT MNSQ

No	INFIT MNSQ Value	Description
1	>1.33	Infit the model
2	0.77-1.33	Fit the model
3	<0.77	Infit the model

The results of the analysis of the INFIT MNSQ value in the QUEST program can be seen in the Figure 1.



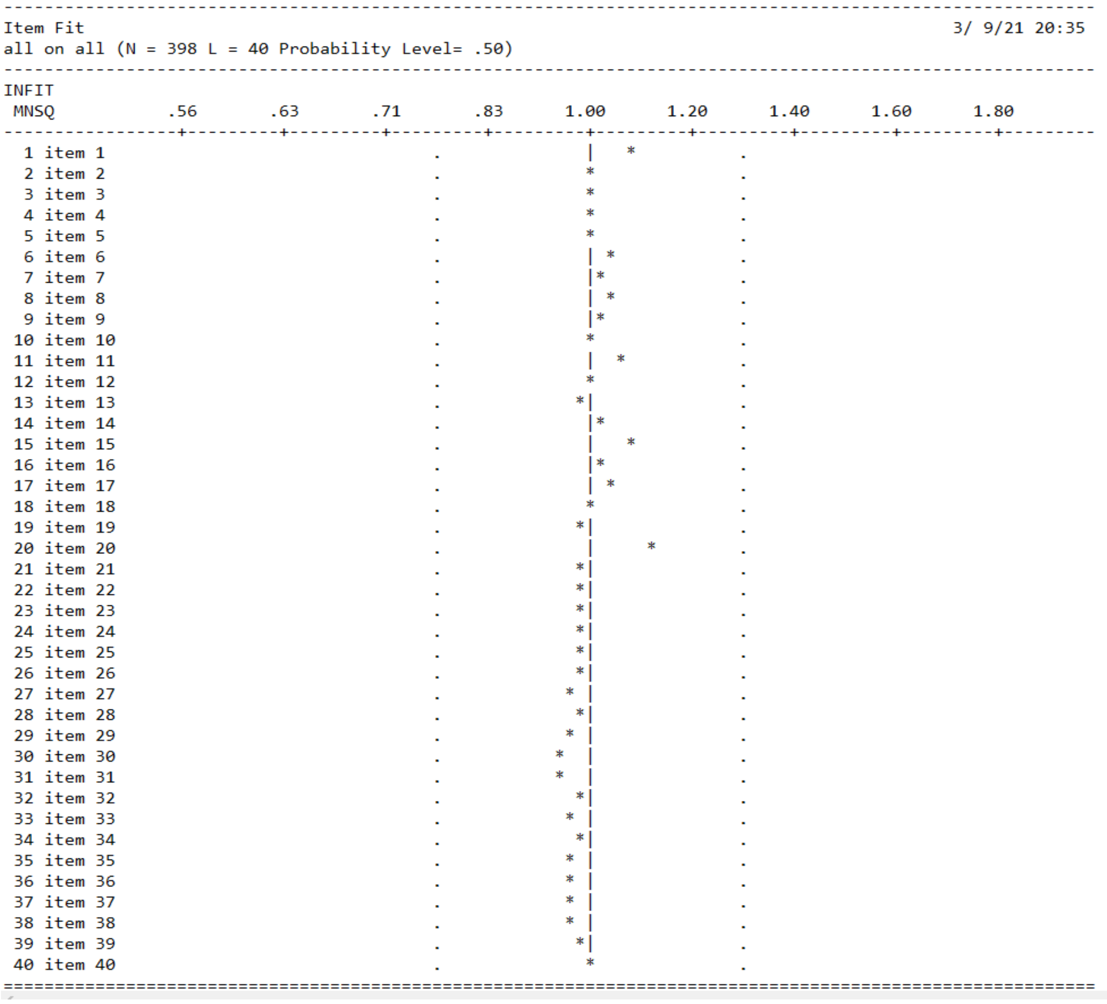
```
------------------------------------------------------------------------
Item Estimates (Thresholds) In input Order
all on all (N = 398 L = 40 Probability Level= .50)
------------------------------------------------------------------------
    ITEM NAME         |SCORE MAXSCR|  THRSH |  INFT  OUTFT INFT  OUTFT
                      |            |    1   |  MNSQ  MNSQ   t     t
------------------------------------------------------------------------
1   item 1            |   124  398 |  -1.46 |  1.09  1.11   2.0   1.6
                      |            |    .11|
                      |            |        |
2   item 2            |     7  398 |   1.74 |  1.00  1.17    .1    .5
                      |            |    .38|
                      |            |        |
3   item 3            |    15  398 |    .98 |  1.01  1.22    .1    .8
                      |            |    .26|
                      |            |        |
4   item 4            |    14  398 |   1.05 |  1.01  1.14    .1    .6
                      |            |    .27|
                      |            |        |
5   item 5            |     6  398 |   1.90 |   .99   .82    .1   -.3
                      |            |    .41|
                      |            |        |
6   item 6            |    87  398 |   -.97 |  1.05  1.07    .7    .7
                      |            |    .12|
                      |            |        |
7   item 7            |    38  398 |    .00 |  1.02  1.04    .2    .3
                      |            |    .17|
                      |            |        |
8   item 8            |    37  398 |    .03 |  1.04  1.22    .3   1.3
                      |            |    .17|
                      |            |        |
9   item 9            |    33  398 |    .15 |  1.02  1.12    .2    .7
                      |            |    .18|
                      |            |        |
10  item 10           |    10  398 |   1.39 |  1.01  1.10    .1    .4
                      |            |    .32|
                      |            |        |
11  item 11           |    96  398 |  -1.10 |  1.07  1.13   1.1   1.4
                      |            |    .12|
                      |            |        |
12  item 12           |    17  398 |    .85 |  1.00  1.05    .1    .3
                      |            |    .25|
                      |            |        |
13  item 13           |    22  398 |    .58 |   .98   .76   -.1  -1.1
                      |            |    .22|
                      |            |        |
14  item 14           |    13  398 |   1.12 |  1.01  1.10    .1    .4
                      |            |    .28|
                      |            |        |
15  item 15           |   121  398 |  -1.42 |  1.09  1.12   1.9   1.6
                      |            |    .11|
                      |            |        |
16  item 16           |    32  398 |    .19 |  1.02  1.16    .2    .9
                      |            |    .19|
                      |            |        |
17  item 17           |    54  398 |   -.39 |  1.04  1.18    .4   1.4
                      |            |    .15|
                      |            |        |
18  item 18           |    13  398 |   1.12 |  1.00   .93    .1   -.1
                      |            |    .28|
                      |            |        |
```

```
------------------------------------------------------------------------
Item Estimates (Thresholds) In input Order
all on all (N = 398 L = 40 Probability Level= .50)
------------------------------------------------------------------------
    ITEM NAME         |SCORE MAXSCR|  THRSH |  INFT  OUTFT INFT  OUTFT
                      |            |    1   |  MNSQ  MNSQ   t     t
------------------------------------------------------------------------
19  item 19           |    31  398 |    .22 |   .99   .96    .0   -.1
                      |            |    .19|
                      |            |        |
20  item 20           |   119  398 |  -1.40 |  1.11  1.17   2.4   2.2
                      |            |    .11|
                      |            |        |
21  item 21           |   116  398 |  -1.36 |   .98   .95   -.5   -.7
                      |            |    .11|
                      |            |        |
22  item 22           |    14  398 |   1.05 |   .99   .81    .0   -.6
                      |            |    .27|
                      |            |        |
23  item 23           |    11  398 |   1.29 |   .99   .91    .1   -.2
                      |            |    .31|
                      |            |        |
24  item 24           |    29  398 |    .29 |   .99   .93    .0   -.3
                      |            |    .19|
                      |            |        |
25  item 25           |    38  398 |    .00 |   .98   .91   -.1   -.5
                      |            |    .17|
                      |            |        |
26  item 26           |    87  398 |   -.97 |   .98   .92   -.3   -.8
                      |            |    .12|
                      |            |        |
27  item 27           |    75  398 |   -.79 |   .96   .90   -.5  -1.0
                      |            |    .13|
                      |            |        |
28  item 28           |    37  398 |    .03 |   .98   .92   -.1   -.5
                      |            |    .17|
                      |            |        |
29  item 29           |    35  398 |    .09 |   .96   .97   -.2   -.1
                      |            |    .18|
                      |            |        |
30  item 30           |    54  398 |   -.39 |   .94   .83   -.6  -1.4
                      |            |    .15|
                      |            |        |
31  item 31           |   100  398 |  -1.15 |   .95   .90   -.9  -1.1
                      |            |    .12|
                      |            |        |
32  item 32           |    45  398 |   -.19 |   .98   .90   -.1   -.6
                      |            |    .16|
                      |            |        |
33  item 33           |    47  398 |   -.23 |   .95   .82   -.4  -1.3
                      |            |    .16|
                      |            |        |
34  item 34           |    29  398 |    .29 |   .99   .92    .0   -.4
                      |            |    .19|
                      |            |        |
35  item 35           |    96  398 |  -1.10 |   .96   .92   -.7   -.9
                      |            |    .12|
                      |            |        |
36  item 36           |    37  398 |    .03 |   .97   .86   -.2   -.9
                      |            |    .17|
                      |            |        |
```

```
------------------------------------------------------------------------
Item Estimates (Thresholds) In input Order
all on all (N = 398 L = 40 Probability Level= .50)
------------------------------------------------------------------------
    ITEM NAME         |SCORE MAXSCR|  THRSH |  INFT  OUTFT INFT  OUTFT
                      |            |    1   |  MNSQ  MNSQ   t     t
------------------------------------------------------------------------
37  item 37           |    55  398 |   -.41 |   .97   .91   -.3   -.7
                      |            |    .15|
                      |            |        |
38  item 38           |    48  398 |   -.26 |   .96   .83   -.4  -1.2
                      |            |    .16|
                      |            |        |
39  item 39           |    30  398 |    .26 |   .97   .86   -.1   -.7
                      |            |    .19|
                      |            |        |
40  item 40           |    92  398 |  -1.04 |  1.00   .99    .1   -.1
                      |            |    .12|
                      |            |        |
------------------------------------------------------------------------
Mean                  |            |    .00 |  1.00   .99    .1    .0
SD                    |            |    .92 |   .04   .13    .7    .9
========================================================================
```

Figure 1. Item validity recapitulation

Figure 1 above provides information about item validity where all items fit or match the Rasch model with an INFIT MNSQ value range between 0.94 – 1.11. To find out if the item fits the Rasch Model, you can also view the item fit map via the Figure 2.



```
-----------------------------------------------------------------------------------------------------------
Item Fit                                                                                    3/ 9/21 20:35
all on all (N = 398 L = 40 Probability Level= .50)
-----------------------------------------------------------------------------------------------------------
INFIT
 MNSQ           .56       .63       .71       .83      1.00      1.20      1.40      1.60      1.80
-----------------+---------+---------+---------+---------+---------+---------+---------+---------+---------
  1 item 1                                .              |   *          .
  2 item 2                                .              *              .
  3 item 3                                .              *              .
  4 item 4                                .              *              .
  5 item 5                                .              *              .
  6 item 6                                .              | *            .
  7 item 7                                .              |*             .
  8 item 8                                .              | *            .
  9 item 9                                .              |*             .
 10 item 10                               .              *              .
 11 item 11                               .              |  *           .
 12 item 12                               .              *              .
 13 item 13                               .             *|              .
 14 item 14                               .              |*             .
 15 item 15                               .              |   *          .
 16 item 16                               .              |*             .
 17 item 17                               .              | *            .
 18 item 18                               .              *              .
 19 item 19                               .             *|              .
 20 item 20                               .              |     *        .
 21 item 21                               .             *|              .
 22 item 22                               .             *|              .
 23 item 23                               .             *|              .
 24 item 24                               .             *|              .
 25 item 25                               .             *|              .
 26 item 26                               .             *|              .
 27 item 27                               .            * |              .
 28 item 28                               .             *|              .
 29 item 29                               .            * |              .
 30 item 30                               .           *  |              .
 31 item 31                               .           *  |              .
 32 item 32                               .             *|              .
 33 item 33                               .            * |              .
 34 item 34                               .             *|              .
 35 item 35                               .            * |              .
 36 item 36                               .            * |              .
 37 item 37                               .            * |              .
 38 item 38                               .            * |              .
 39 item 39                               .             *|              .
 40 item 40                               .              *              .
===========================================================================================================
```

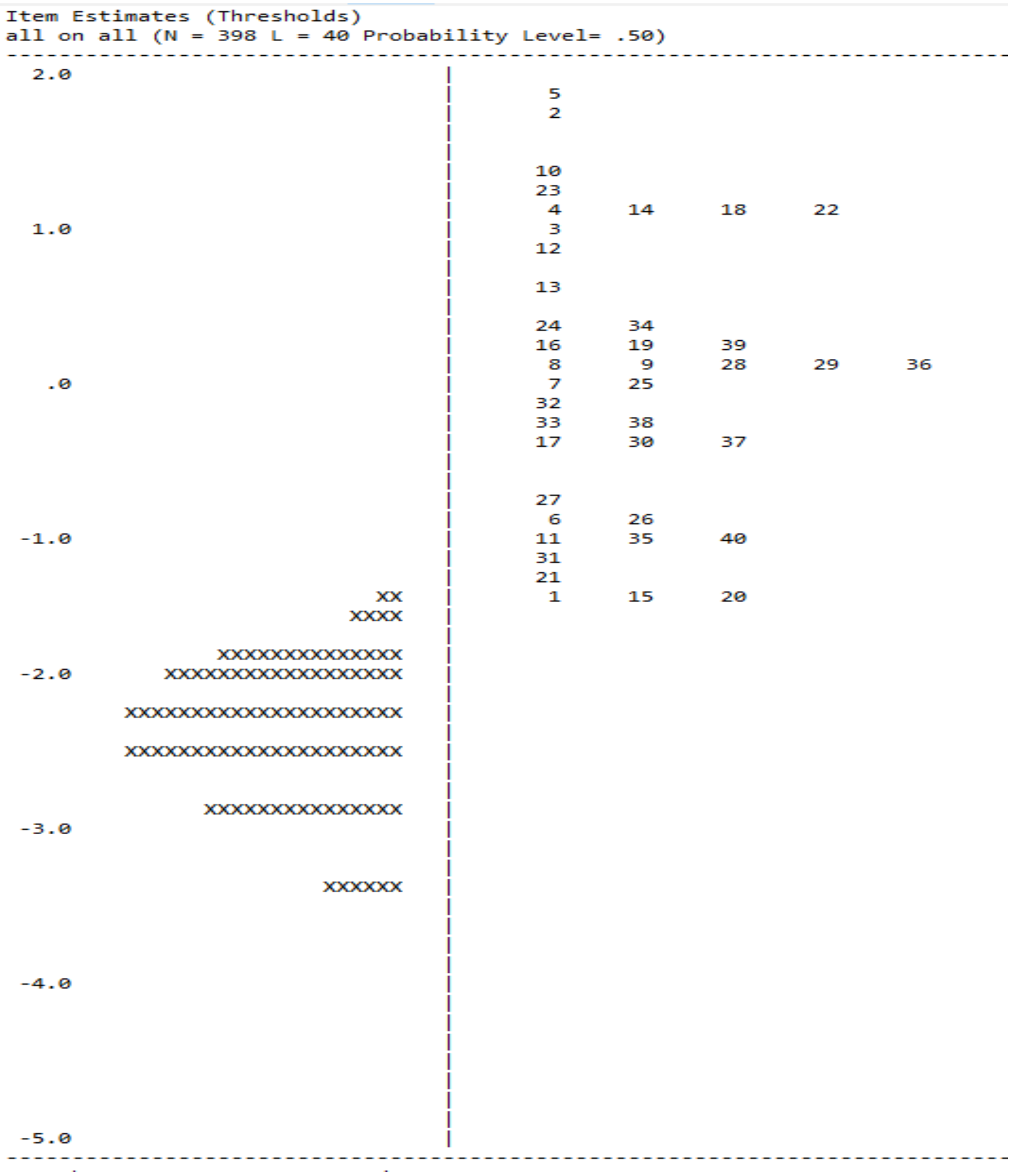
Figure 2. Rasch Model Fit Map

When viewed from the fit map of the model above, it is known that all items are in the INFIT MNSQ value range 0.77 – 1.30. The dots on the left show the value 0.77 while the dots on the right show the value 1.30.

Difficulty Estimation

To find out the difficulty level of an item through the QUEST program, it can be seen by looking at the results of the item estimate (Threshold) analysis. The criteria for determining the difficulty level of an item revolve around the value of -2.0 – 2.0. If the range or distribution of items or test takers < -2.0, then the item is included in the easy category. Meanwhile, if the range or distribution of items or test takers >2.0, then the item is included in the difficult category. For a more detailed view of the distribution of item difficulty levels, see the Figure 3.

In Figure 3, item number 5 is the most difficult item. Even if compared to the ability of the test takers, the possibility of the test taker correctly answering item number 5 is very small or it can be said that it is impossible.



```
Item Estimates (Thresholds)
all on all (N = 398 L = 40 Probability Level= .50)
----------------------------------------------------------------------------
  2.0                            |
                                 |       5
                                 |       2
                                 |
                                 |
                                 |      10
                                 |      23
                                 |       4     14     18     22
  1.0                            |       3
                                 |      12
                                 |
                                 |      13
                                 |
                                 |      24     34
                                 |      16     19     39
                                 |       8      9     28     29     36
   .0                            |       7     25
                                 |      32
                                 |      33     38
                                 |      17     30     37
                                 |
                                 |
                                 |      27
                                 |       6     26
 -1.0                            |      11     35     40
                                 |      31
                                 |      21
                            XX   |       1     15     20
                          XXXX   |
                                 |
                XXXXXXXXXXXXXX   |
 -2.0       XXXXXXXXXXXXXXXXXX   |
                                 |
         XXXXXXXXXXXXXXXXXXXXX   |
                                 |
         XXXXXXXXXXXXXXXXXXXXX   |
                                 |
                                 |
               XXXXXXXXXXXXXXX   |
 -3.0                            |
                                 |
                                 |
                        XXXXXX   |
                                 |
                                 |
                                 |
                                 |
 -4.0                            |
                                 |
                                 |
                                 |
                                 |
                                 |
                                 |
                                 |
 -5.0                            |
----------------------------------------------------------------------------
```

Figure 3. Distribution of Item Difficulty

In addition, item number 20 is the easiest item and is in accordance with the ability of the test taker. The level of item difficulty through the QUEST program can also be seen from the item estimate threshold with the criteria in Table 2.

Table 2. Criteria for Item Difficulty

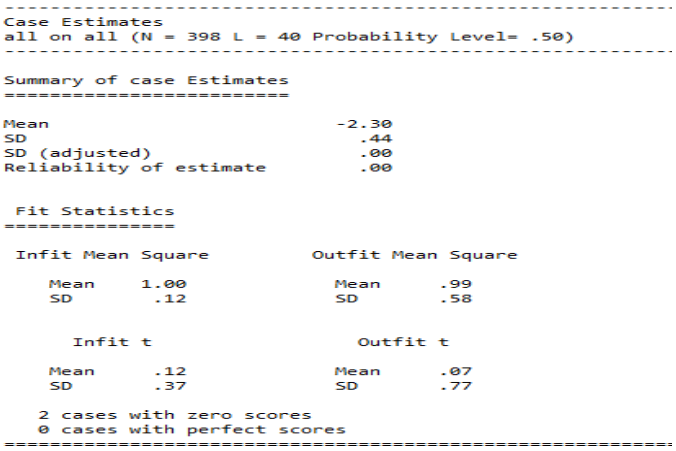
No	Threshold Value	Description
1	$b > 2$	Very difficult
2	$1 < b \leq 2$	Difficult
3	$-1 \leq b \leq 1$	Medium
4	$-2 \leq b < -1$	Easy
5	$b < -2$	Very easy

Table 3 presents a recapitulation of the difficulty level of each item.

Table 3. Recapitulation of the Difficulty Level of the Rasch Model Questions

Item	Threshold Value	Description	Item	Threshold Value	Description
1	-1.46	Easy	21	-1.36	Easy
2	1.74	Difficult	22	1.05	Difficult
3	0.98	Medium	23	1.29	Difficult
4	1.05	Difficult	24	0.29	Medium
5	1.90	Difficult	25	0.00	Medium
6	-0.97	Medium	26	-0.97	Medium
7	0.00	-	27	-0.79	Medium
8	0.03	Medium	28	0.03	Medium
9	0.15	Medium	29	0.09	Medium
10	1.39	Difficult	30	-0.39	Medium
11	-1.10	Easy	31	-1.15	Easy
12	0.85	Medium	32	-0.19	Medium
13	0.58	Medium	33	-0.23	Medium
14	1.12	Difficult	34	0.29	Medium
15	-1.42	Easy	35	-1.10	Easy
16	0.19	Medium	36	0.03	Medium
17	-0.39	Medium	37	-0.41	Medium
18	1.12	Difficult	38	-0.26	Medium
19	0.22	Medium	39	0.26	Medium
20	-1.40	Easy	40	-1.04	Easy

The level of difficulty based on Table 3 can be illustrated that the item in the very difficult category is 0%. Items in the difficult category are 8 items or 21%, the medium item category is 23 items or 59%, the easy category is 8 items or 21%. In general, the test taker's ability is below the item difficulty level. This is evidenced by the small number of test participants who are able to correctly answer items with difficulty. To determine the ability of the test takers through the QUEST program, see the Summary of Case Estimate with criteria, if the Estimate value is > 1.00 in the high ability category, -1.00 – 1.00 moderate ability, and < -1.00 low ability in Figure 4.



```
-----------------------------------------------------------
Case Estimates
all on all (N = 398 L = 40 Probability Level= .50)
-----------------------------------------------------------

Summary of case Estimates
=========================

Mean                             -2.30
SD                                 .44
SD (adjusted)                      .00
Reliability of estimate            .00


 Fit Statistics
 ===============

 Infit Mean Square          Outfit Mean Square

    Mean     1.00             Mean      .99
    SD        .12             SD        .58


      Infit t                   Outfit t

    Mean      .12             Mean      .07
    SD        .37             SD        .77

   2 cases with zero scores
   0 cases with perfect scores
===========================================================
```

Figure 4. Estimation of Respondents' Ability

Figure 4 provides information that the test takers have moderate ability, with a reliability estimate value of 0.00 or with a range of -1.00 – 1.00.

Estimated Item Fit

To find out which items fall or pass based on the OUTFIT t value in the QUEST program. If value OUTFIT t ≥ 2.00 fall items in Table 4.

Table 4. Fit Item Recapitulation

Item	Outfit t Value	Description	Item	Outfit t Value	Description
1	1.6	Fit	21	-0.7	Fit
2	0.5	Fit	22	-0.6	Fit
3	0.8	Fit	23	-0.2	Fit
4	0.6	Fit	24	-0.3	Fit
5	-0.3	Fit	25	-0.5	Fit
6	0.7	Fit	26	-0.8	Fit
7	0.3	Fit	27	-1.0	Fit
8	1.3	Fit	28	-0.5	Fit
9	0.7	Fit	29	-0.1	Fit
10	0.4	Fit	30	-1.4	Fit
11	1.4	Fit	31	-1.1	Fit
12	0.3	Fit	32	-0.6	Fit
13	-1.1	Fit	33	-1.3	Fit
14	0.4	Fit	34	-0.4	Fit
15	1.6	Fit	35	-0.9	Fit
16	0.9	Fit	36	-0.9	Fit
17	1.4	Fit	37	-0.7	Fit
18	-0.1	Fit	38	-1.2	Fit
19	-0.1	Fit	39	-0.7	Fit
20	2.2	Infit	40	-0.1	Fit

Based on Table 4, 39 items passed so that they could be used and there was 1 item that did not pass so that it could not be used. The items used in the final semester exam have a proportional level of difficulty (Mardapi, 2017), so that the questions that have been analyzed meet the ideal criteria to be used as formative tests or end-of-semester exams. The level of suitability of this item is used to see the accuracy of the item with the model or item fit. The level of conformity of the goods describes whether our goods function normally to take measurements or not. If there are items that are not appropriate, this indicates a subject's misconception in answering the questions (Camminatiello et al., 2010).

Estimated Reliability

The reliability value of the Rasch model using the QUEST program is seen in the reliability of item estimate and reliability of case estimate. In the reliability of item estimate value of 0.95. In Rasch modeling, this reliability is referred to as sample reliability. The criteria for the reliability value of the Rasch model as stated in the opinion (Susdelina et al., 2018) as follows; <0.67 low, 0.67-0.80 enough, 0.81 – 0.90

good, 0.91 – 0.94 very good, >0.94 perfect. The reliability of item estimate value of 0.95 is related to the number of items that fit the model.

The value of 0.94 includes reliability with a very good category so that it affects the items that fit the model. The higher the reliability, the more items fit with the model. While the reliability of case estimate value or the reliability of test participants of 0.00 is classified as low. This value indicates that there is an inconsistency as expressed (Ardiyanti, 2016) In the test taker's answer, the inconsistency of the test taker's answer can also mean that the test taker is careless in answering the questions so that it affects the reliability value of the person/subject to be low. (Pratama, 2020).

CONCLUSION

Based on the results of the analysis of the Final Semester Exams, several characteristics of the test and test takers can be described as follows; 1) the estimated validity of items fit or matched the Rasch model for 40 items with an INFIT MNSQ value range between 0.94 – 1.11 and all items on the test can be used based on the results of the estimated OUTFIT t value ≤ 2.00, OUTFIT t analysis obtained 39 items that fit. 2) all items were analyzed with the estimated level of difficulty of the items in the very difficult category of 0%. Items in the difficult category are 8 items or 21%, the medium item category is 23 items or 59%, the easy category is 8 items or 21% and the items in the very easy category are 0%. In general, the test taker's ability is below the item difficulty level. 3) the value of reliability of item estimate is 0.95 with very good category and the value of reliability of case estimate is 0.00 with weak category. Based on the results of the analysis using the Rasch model, in general, the instrument for this semester's final exam can be used. However, it is not appropriate if the measurement results are used for decision making based on students' abilities (Primi et al., 2016).

REFERENCES

Abdullah, H., Arsad, N., Hashim, F. H., Aziz, N. A., Amin, N., & Ali, S. H. (2012). Evaluation of Students' Achievement in the Final Exam Questions for Microelectronic (KKKL3054) using the Rasch Model. *Procedia Social and Behavioral Sciences*, *60*(c), 119–123. https://doi.org/10.1016/j.sbspro.2012.09.356

Alkharusi, H. (2015). An evaluation of the measurement of perceived classroom assessment environment. *International Journal of Instruction*, *8*(2), 45–54. https://doi.org/10.12973/iji.2015.824a

Anwar, M. S., Choirudin, C., Ningsih, E. F., Dewi, T., & Maseleno, A. (2019). Developing an Interactive Mathematics Multimedia Learning Based on Ispring Presenter in Increasing Students' Interest in Learning Mathematics. *Al-Jabar : Jurnal Pendidikan Matematika*, *10*(1), 135–150. https://doi.org/10.24042/ajpm.v10i1.4445

Ardiyanti, D. (2016). Aplikasi Model Rasch pada Pengembangan Skala Efikasi Diri dalam Pengambilan Keputusan Karir Siswa. *Jurnal Psikologi*, *43*(3), 248–263.

Aricak, O. T., Avcu, A., Topcu, F., & Tutlu, M. G. (2020). Use of Item Response Theory to Validate Cyberbullying Sensibility Scale for University Students. *International Journal of Assessment Tools in Education*, *7*(1), 18–29. https://doi.org/10.21449/ijate.629584

Camminatiello, I., Gallo, M., & Menini, T. (2010). The Rasch Model for Evaluating Italian Student Performance. *Journal of Applied Quantitative Methods*, *5*(2), 331–349.

Clark, I. (2012). Formative Assessment : Assessment Is for Self-regulated Learning. *JSTOR*, *24*(2), 205–249. https://doi.org/10.1007/S10648-01

Djemari, M. (2012). *Pengukran Penilaian dan Evaluassi Pendidikan*. Nuha Medika.

Fitriani, L., Ramalis, T. R., & Efendi, R. (2019). Karakterisasi Tes Keterampilan Proses Sains Materi Fluida Statis Berdasarkan Teori Respon Butir. *Omega: Jurnal Fisika dan Pendidikan Fisika*, *5*(2), 27. https://doi.org/10.31758/omegajphysphyseduc.v5i2.27

Gipps, C., Hargreaves, E., & McCallum, B. (2015). *What makes a good primary school teacher? Expert classroom strategies*. Routledge.

Griffrin, P., & McGaw, B. (2014). Assesment and Teaching of 21st Century Skills. In *Assessment and teaching of 21st century skills*. https://doi.org/10.1007/978-94-007-2324-5_2

Hayati, S., & Lailatussaadah, L. (2016). Validitas Dan Reliabilitas Instrumen Pengetahuan Pembelajaran Aktif, Kreatif Dan Menyenangkan (Pakem) Menggunakan Model Rasch. *Jurnal Ilmiah Didaktika*, *16*(2), 169. https://doi.org/10.22373/jid.v16i2.593

Imaroh, N., Susongko, P., & Isnani. (2017). Uji Validitas Tes Ulangan Akhir Semester Gasal Mata Pelajaran Matematika. *E-Journal UPS*, *1*(1), 80–89.

Kadir, A. (2015). Menyusun dan Menganalisis Tes Hasil Belajar. *Al-Ta'dib*, *8*(2), 70–81.

Kartianom, K., & Ndayizeye, O. (2017). What 's wrong with the Asian and African Students' mathematics learning achievement? The multilevel PISA 2015 data analysis for Indonesia, Japan, and Algeria. *Jurnal Riset Pendidikan Matematika*, *4*(2), 200. https://doi.org/10.21831/jrpm.v4i2.16931

Kartianom, & Mardapi, D. (2017). The utilization of junior high school mathematics national examination data: A conceptual error diagnosis. *REiD (Research and Evaluation in Education)*, *3*(2), 163–173. https://doi.org/10.1044/2018_AJSLP-17-0074

Marjiastuti, K., & Wahyuni, S. (2014). Analisis kemampuan peserta didik dengan model Rasch. In *Seminar Nasional Evaluasi Pendidikan II*.

Naqiyah, M., Rosana, D., Sukardiyono, & Ernasari. (2020). *Developing Instruments to Measure Physics Problem Solving Ability and Nationalism of High School Student*. *13*(4), 921–936.

OECD. (2005). *Formative Assessment: Improving Learning in Secondary Classrooms*. *29*(November), 282. http://new.sourceoecd.org/9264007393

Pey Tee, O., & Subramaniam, R. (2018). Comparative study of middle school students' attitudes towards science: Rasch analysis of entire TIMSS 2011 attitudinal data for England, Singapore and the U.S.A. as well as psychometric properties of attitudes scale. *International Journal of Science Education*, *40*(3), 268–290. https://doi.org/10.1080/09500693.2017.1413717

Pratama, D. (2020). Analisis Kualitas Tes Buatan Guru Melalui Pendekatan Item Response Theory (IRT) Model Rasch. *Tarbawy : Jurnal Pendidikan Islam*, *7*(1), 61–70. https://doi.org/10.32923/tarbawy.v7i1.1187

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The Development and Testing of a New Version of the Cognitive Reflection Test Applying Item Response Theory (IRT). *Journal of Behavioral Decision Making*, *29*(5), 453–469. https://doi.org/10.1002/bdm.1883

Purwanto. (2009). *Evaluasi Hasil Belajar*. Pustaka Pelajar.

Rahayuningsih, S., & Jayanti, R. (2019). High Order Thinking Skills (HOTS) Students In Solving Group Problem Based Gender. *Al-Jabar : Jurnal Pendidikan Matematika*, *10*(2), 243–250. https://doi.org/10.24042/ajpm.v10i2.4872

Retnawati, Hari. (2014). *Teori respons butir dan penerapannya : Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Nuha Mediak.

Retnawati, H., & Hadi, S. (2014). Sistem bank soal daerah terkalibrasi untuk menyongsong era desentralisasi. *Jurnal Ilmu Pendidikan*, *20*(2), 183–193.

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. https://doi.org/10.12973/iji.2016.914a

Rindermann, H., & Baumeister, A. E. E. (2015). Validating the Interpretations of PISA and TIMSS Tasks: A Rating Study. *International Journal of Testing*, *15*(1), 1–22. https://doi.org/10.1080/15305058.2014.966911

Safihin, M. (2019). Pengembangan Tes Menggunakan Model Rasch Materi Gaya Untuk SMA. *Jurnal Pendidikan dan Pembelajaran*, *8*(6), 1–11. http://jurnal.untan.ac.id/index.php/jpdpb/article/view/33424/75676581548

Santoso, A., Kartianom, K., & Kassymova, G. K. (2019). Kualitas butir bank soal statistika (Studi kasus: Instrumen ujian akhir mata kuliah statistika Universitas Terbuka). *Jurnal Riset Pendidikan Matematika*, *6*(2), 165–176. https://doi.org/10.21831/jrpm.v6i2.28900

Shute, V. J., & Becker, B. J. (2010). Innovative Assessment for the 21st Century. *Innovative Assessment for the 21st Century*. https://doi.org/10.1007/978-1-4419-6530-1

Siti Aminah, N. (2013). Jurnal Materi dan Pembelajaran Fisika (JMPF). *Jurnal Materi dan Pembelajaran Fisika (JMPF)*, *3*(2), 33–39.

Sumintono, B. (2014). Model Rasch untuk Penelitian Sosial Kuantitatif. Institut Teknologi Sepuluh November (Nomor November 201). http://deceng3.wordpress.com

Susanto, H., Rinaldi, A., & Novalia. (2015). Analisis Validitas Reabilitas Tingkat Kesukaran dan Daya Beda pada Butir Soal Ujian Akhir Semester Ganjil Mata Pelajaran Matematika. *Al-Jabar: Jurnal Pendidikan Matematika*, *6*(2), 203–217. https://doi.org/10.18907/jjsre.37.3_343_4

Susdelina, Perdana, S. A., & Febrian. (2018). Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik Dan Rasch Model. *Jurnal Kiprah*, *6*(1), 41–48. https://doi.org/10.31629/kiprah.v6i1.574

Tri Wahyuningsih, E. (2015). *Analisis Butir Soal Tes Objektif Buatan Guru Ulangan Semester Ganjil Mata Pelajaran Ekonomi Kelas X di Sma Negeri 1 Mlati Tahun Ajaran 2013/2014*. Universitas Negeri Yogyakarta.

Wafubwa, R. N., & Csíkos, C. (2020). Formative Assessment as a Predictor of Mathematics Teachers' Levels of Metacognitive Regulation. *International Journal of Instruction*, *14*(1), 983–998. https://doi.org/10.29333/IJI.2021.14158A

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Educational Measurement for Applied Researchers. In *Educational Measurement for Applied Researchers*. https://doi.org/10.1007/978-981-10-3302-5