



Implementation of Item Response Theory (IRT) Rasch Model in Quality Analysis of Final Exam Tests in Mathematics

Abdul Rahim[✉], Haryanto Haryanto

Universitas Negeri Yogyakarta, Indonesia

Article Info

History Articles

Received:

13 May 2021

Accepted:

07 June 2021

Published:

30 August 2021

Keywords:

Fit, rasch model, winsteps, reliability

Abstract

Item Response Theory IRT is very useful in item analysis of a test. Through IRT, the index of item parameters can be known easily. The index becomes the basis for selecting items. In addition, the information function can give consideration to how the test should be used. This study was conducted to analyze the test instrument used to measure the ability of students in the odd semester final exam in mathematics. Sampling using purposive sampling technique. These students consist of 67 people. The questions given are in the form of multiple choice questions totaling 40 items related to the odd semester final exam material. The data analysis technique used quantitative descriptive analysis. The Rasch model is used to obtain fit items. This analysis was carried out with the help of Winsteps 3.73 software. From the output of the Winsteps program, 35 items were obtained according to the Rasch model with an average value of Outfit MNSQ for persons and items of 1.09 and 1.09, respectively. While the Outfit ZSTD values for person and item are -0,1 and -0,2 respectively. Meanwhile, the instrument reliability expressed in Cronbach's alpha is 0.77.

[✉] Correspondence Address :

Department of Educational Research and Evaluation

Graduate School

Universitas Negeri Yogyakarta, Indonesia 55281

E-mail: abdulrahim.2019@student.uny.ac.id

p-ISSN 2252-6420

e-ISSN 2503-1732

INTRODUCTION

In learning in the school environment, teachers play an important role in evaluating and assessing the subjects being taught (Palinussa & Thaib, 2020; Sutama et al., 2017). Assessment conducted by the teacher to see progress (Santoso, 2014), improve the learning process (Umami, 2018) and obtain information about students by giving assignments, written tests, class questions and answers, daily tests, mid-semester tests, and end-of-semester tests. (R. Abdullah, 2017; Nurjannah, 2017). The evaluation process needs to be carried out properly in order to be able to measure the actual abilities of students (B, 2017; Talakua et al., 2020). A good assessment requires a good test instrument too (Agustin et al., 2018; Nufus et al., 2017). A good instrument is an instrument that has a valid measuring power (Muluki et al., 2020). Instruments that do not have valid measuring power will not be able to provide any information regarding the test taker's ability (Solichin, 2017). In fact, many test instruments still do not know the quality of the questions, so there is a quasi-assessment which has an impact on the actual ability of students who cannot be measured (Fauziana & Wulansari, 2021).

Item analysis can help improve the quality of questions through revision or discarding ineffective questions, besides that it can be used as diagnostic information for students, whether they have understood the material (Muharromah & Humaisi, 2020). In the field of education, there are two approaches used to analyze test quality, namely Classical Test Theory (CTT) and Item Response Theory (IRT) (Fitriani et al., 2019). However, analysis of test quality using the classical test approach has been abandoned because it has many weaknesses (Pratama, 2020). Classical test theory has several weaknesses, namely; the results of the measurement will depend on the characteristics of the test used, the parameter of the item depends on the ability of the test taker, and measurement error can only be

known for groups, not individuals (Mardapi, 2012).

In contrast to the Item Response Theory (IRT) approach, it is a general statistical theory about examining items and testing performance and how performance relates to abilities as measured by items in the test (Aprita & Haryati, 2021). IRT is one way to assess the feasibility of items by comparing and appearance of items against the appearance of evidence of group ability predicted by the model (van der Linden & Hambelton, 2013). According to (Liang et al., 2014) say that "*Item response theory (IRT) is a powerful scaling technique with appealing features such as the invariance of item and ability parameter values*". IRT assumes that the probability of a test taker answering correctly for each item depends on the ability of the test taker. Thus, test takers who have high abilities have a greater chance of answering correctly than those with low abilities (Retnawati, 2014). According to (Hambelton & Swaminathan, 2013) IRT has several advantages, namely; The score describes the test taker's ability and does not depend on the difficulty of the test, can be used to relate the item to the test taker's ability, and does not require parallel tests to determine the reliability coefficient.

One of the simplest IRT models that has been widely used by experts in developing tests is the Rasch model, with one parameter (1-PL) (Chan et al., 2021; Falani & Kumala, 2017; Sumaryanto & Khumaedi, 2019). Rasch model is very easy to do and apply with accurate analysis results (Che Lah et al., 2021; Susdelina et al., 2018), also reviewing the opportunity to answer correctly on the questions by comparing students' abilities with the level of difficulty of the questions (Nielsen et al., 2021; Sumintono & Widhiarso, 2015). The Rasch model has a variable difficulty level, regardless of the sample involved in the initial variation (Scoulas et al., 2021; Wei et al., 2012). Rasch developed a data measurement model that can determine the relationship between the student's own level of ability (person ability) and the level of item difficulty by using a logarithmic function to be

able to produce measurements with the same interval value (Bambang, 2014; Tseng & Wang, 2021). The main characteristic of the Rasch model is that it considers all responses from a test taker regardless of the order of problem solving (Isnani et al., 2019). In addition, the selection of the Rasch model is because this model at least meets the principles of the 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 (H. Abdullah et al., 2012; Sumintono, 2014).

The relevance of this research is the research conducted by (Alfarisa & Purnama, 2019) shows that the analysis using the Rasch model approach can explain the quality of the test items. Other research on item analysis using Rasch modeling can explain an item and the abilities of students in permutations and combinations of mathematics learning (Dwinata, 2019). Furthermore, other research conducted by (Imaroh et al., 2020) analyzing the items using the Rasch model can provide information about the quality of the items in the final test of the odd semester mathematics class VII Junior High School (SMP).

Based on the previous research in this study, the researcher wanted to know the quality of the test instrument used to measure students' abilities in the odd semester final exam in mathematics for class VIII SMP with the Rasch model approach. This quality is measured based on several indicators, namely items that fit the Rasch model, level of item difficulty, and item reliability. Therefore, a test instrument was designed and then determined which items were fit and which did not fit the Rasch model. In addition, Cronbach's alpha value will be determined to determine the reliability of the items.

METHODS

This research is focused on the analysis of the final semester examination test

instrument using 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 67 people. There are 40 multiple choice questions with a correct score of 1 and an incorrect score of 0 on the final semester exam test instrument aimed at students. The technique is the data analysis used is descriptive quantitative. The test results in the form of dichotomous scores were analyzed using Winsteps software version 3.73. From the output of the Winsteps software version 3.73, several parameter items were obtained that fit the Rasch model. In addition, Cronbach's alpha value is obtained which is the result of the overall item reliability test. While the MNSQ Outfit, ZSTD Outfit and the correlation value of the item with the question as a whole show the limit of items that are declared fit with the model. That is, if the Outfit MNSQ value is between 0.5 to 1.5; Outfit ZSTD value is between -2.0 to 2.0; and the correlation value of the item with the total score is positive (Sumintono & Widhiarso, 2015).

RESULTS AND DISCUSSION

Summary of Statistics

Based on data analysis using Winsteps 3.73 software, there are 35 items that fit the Rasch model and 5 other items do not fit the Rasch model. These results are fully presented in Table 1.

Table 1. Summary statistics

	Description	Score
Logit	Person	-0.84
	Item	0.00
Reliabilitas	Person Reliability	0.72
	Item Reliability	0.87
	Alpha Cronbach	0.77
Outfit MNSQ	Person	1.08
	Item	1.08
Outfit ZSTD	Person	-0.1
	Item	-0.2

Table 1 shows the logit value of the person or measure measure of -0.84 and the item measure value of 0 which means the person measure value is smaller than the item measure. It can be stated that the ability of students tends to be lower than the level of difficulty of the questions. Meanwhile, Person Reliability is worth 0.72, Item reliability is 0.87, and Cronbach's Alpha value is 0.77. From this value, it can be stated that the level of consistency of answers from students is good, namely 0.72, and the quality of the items on the test instrument used has very good reliability, which is 0.87. In addition, the value of Cronbach's Alpha which shows the interaction between person and item as a whole has a good value of 0.77.

Another quantity shown in table 1 is the Outfit Mean Squared (Outfit MNSQ) value of 1.08 in both the person and item columns. The value of 1.08 is included in the fit criteria, which is between 0.5 <MNSQ <1.5, meaning that the test instrument used is in accordance with the model to measure the competence of

students in the final semester exam. Furthermore, the Outfit Z Standardized value (Outfit ZSTD) is -0.1 for the person and -0.2 for the item. The values of -0.1 and -0.2 are between -2.0 <ZSTD < 2.0, which means that the data has a possible rational value. This means that overall the questions or items are in accordance with the Rasch model and can be used as an achievement test instrument in the final semester exam.

Level Item Fit

The distribution of item items that are considered to be misfit or not fit with the model can be seen in figure 1. The item limit is declared fit with the model if it meets one or both of the following conditions. The first requirement is that the Outfit MNSQ value is between 0.5 and 1.5; Outfit ZSTD value is between -2.0 to 2.0; and the item correlation value with the total score (point measure correlation) is positive (Sumintono & Widhiarso, 2015).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL		INFIT		OUTFIT		PT-MEASURE		EXACT OBS%	MATCH EXP%	ITEM
				S.E.	MNSQ	ZSTD	MNSQ	ZSTD	CORR.	EXP.				
36	7	67	1.52	.41	1.18	.7	1.93	1.9	A-.16	.22	89.6	89.5	Butir_36	
1	19	67	.20	.29	1.29	2.1	1.55	2.7	B-.10	.32	67.2	72.5	Butir_1	
39	9	67	1.22	.37	1.21	.9	1.54	1.4	C-.10	.25	86.6	86.5	Butir_39	
37	18	67	.28	.29	1.35	2.4	1.49	2.3	D-.15	.32	67.2	73.6	Butir_37	
31	13	67	.75	.32	1.13	.7	1.43	1.5	E .06	.29	80.6	80.6	Butir_31	
32	8	67	1.37	.39	1.13	.6	1.41	1.1	F .01	.24	88.1	88.0	Butir_32	
33	13	67	.75	.32	1.16	.9	1.39	1.4	G .03	.29	80.6	80.6	Butir_33	
29	14	67	.65	.31	1.21	1.2	1.37	1.5	H-.01	.29	77.6	79.1	Butir_29	
24	9	67	1.22	.37	1.05	.3	1.32	.9	I .12	.25	86.6	86.5	Butir_24	
28	12	67	.85	.33	1.13	.7	1.32	1.1	J .07	.28	82.1	82.1	Butir_28	
40	12	67	.85	.33	1.03	.2	1.32	1.1	K .19	.28	82.1	82.1	Butir_40	
34	17	67	.37	.30	1.17	1.2	1.30	1.4	L .07	.31	76.1	74.9	Butir_34	
27	7	67	1.52	.41	1.02	.2	1.28	.8	M .14	.22	89.6	89.5	Butir_27	
35	20	67	.12	.28	1.11	.9	1.27	1.5	N .16	.33	68.7	71.4	Butir_35	
38	16	67	.46	.30	1.15	1.0	1.26	1.2	O .09	.31	77.6	76.3	Butir_38	
26	16	67	.46	.30	1.16	1.1	1.22	1.0	P .09	.31	74.6	76.3	Butir_26	
25	14	67	.65	.31	1.16	.9	1.22	.9	Q .09	.29	77.6	79.1	Butir_25	
30	15	67	.55	.31	1.13	.8	1.14	.7	R .14	.30	76.1	77.7	Butir_30	
2	26	67	-.33	.27	1.08	.8	1.08	.7	S .25	.34	59.7	67.9	Butir_2	
23	16	67	.46	.30	1.01	.1	1.05	.3	T .28	.31	77.6	76.3	Butir_23	
3	48	67	-1.89	.28	1.02	.2	1.04	.3	t .27	.30	70.1	72.2	Butir_3	
21	27	67	-.40	.27	1.01	.2	1.02	.2	s .33	.35	68.7	67.7	Butir_21	
17	25	67	-.26	.27	1.02	.2	1.00	.0	r .33	.34	67.2	68.2	Butir_17	
18	24	67	-.19	.27	1.00	.0	.96	-.2	q .35	.34	67.2	68.6	Butir_18	
6	25	67	-.26	.27	.97	-.2	.96	-.2	p .38	.34	67.2	68.2	Butir_6	
19	29	67	-.54	.26	.93	-.7	.94	-.6	o .43	.35	71.6	67.2	Butir_19	
7	31	67	-.68	.26	.92	-.9	.90	-1.0	n .45	.35	68.7	66.6	Butir_7	
5	35	67	-.95	.26	.92	-1.0	.91	-.9	m .45	.35	70.1	65.3	Butir_5	
4	18	67	.28	.29	.91	-.6	.90	-.5	l .42	.32	76.1	73.6	Butir_4	
8	21	67	.04	.28	.89	-.9	.86	-.9	k .47	.33	73.1	70.4	Butir_8	
20	24	67	-.19	.27	.87	-1.2	.86	-1.1	j .50	.34	73.1	68.6	Butir_20	
10	45	67	-1.66	.27	.86	-1.4	.75	-1.8	i .52	.32	71.6	68.6	Butir_10	
22	24	67	-.19	.27	.83	-1.7	.78	-1.7	h .56	.34	76.1	68.6	Butir_22	
11	42	67	-1.44	.27	.79	-2.4	.71	-2.5	g .61	.33	74.6	65.8	Butir_11	
16	36	67	-1.02	.26	.76	-3.0	.72	-3.0	f .65	.34	79.1	65.1	Butir_16	
12	40	67	-1.30	.26	.74	-3.2	.68	-3.1	e .67	.34	74.6	65.0	Butir_12	
9	34	67	-.88	.26	.71	-3.7	.69	-3.6	d .71	.35	83.6	65.6	Butir_9	
15	31	67	-.68	.26	.70	-3.8	.68	-3.6	c .72	.35	89.6	66.6	Butir_15	
14	35	67	-.95	.26	.67	-4.4	.64	-4.2	b .76	.35	85.1	65.3	Butir_14	
13	32	67	-.75	.26	.62	-5.0	.61	-4.6	a .81	.35	91.0	66.3	Butir_13	
MEAN	22.7	67.0	.00	.30	1.00	-.4	1.09	-.2			76.6	73.6		
S.D.	10.8	.0	.87	.04	.18	1.7	.30	1.9			7.7	7.5		

Figure 1. Data on the Distribution of Misfit or Not Fit Questions with the Rasch model

Based on the results of the analysis of the achievement test instrument using the Winsteps version 3.73 program which is in figure 1, it is obtained that there are 5 items that are misfit, namely item 1, item 29, item 36, item 37, and item 39, and fit questions are 35 items, so obtained the final instrument as many as 35 items.

Person Map Item

Rasch analysis has a feature that uses winstep is that there is a map that describes the distribution of the subject's ability and the distribution of item difficulty levels with the same scale. This map is called the Wright Map which is nothing but a person-item map (Salman & Abd.aziz, 2015).

Based on figure 2 in the left side is the distribution of subject abilities, while on the right side is the distribution of items. From the map, it can be seen that in general the questions in the test are more difficult than the subject's ability. The most difficult items are item 27 and item 36 which are in the topmost position. Theoretically with that question, there will be no subject who has the opportunity to answer the question correctly because it has a lower ability than the level of difficulty of the question. To see the difficulty level of the item in more detail, it will be reviewed below.

Item Difficulty Level

One thing we need to pay attention to is the results of Rasch's analysis with this winsteps. A high logit (measure) value indicates that the item has a high level of difficulty. It correlates with the total score, where multiple correct answers in the total score correlate with a higher measure score. This data size also has the same scale. To find out the classification of the level of difficulty of each item, it can be seen in table 2.

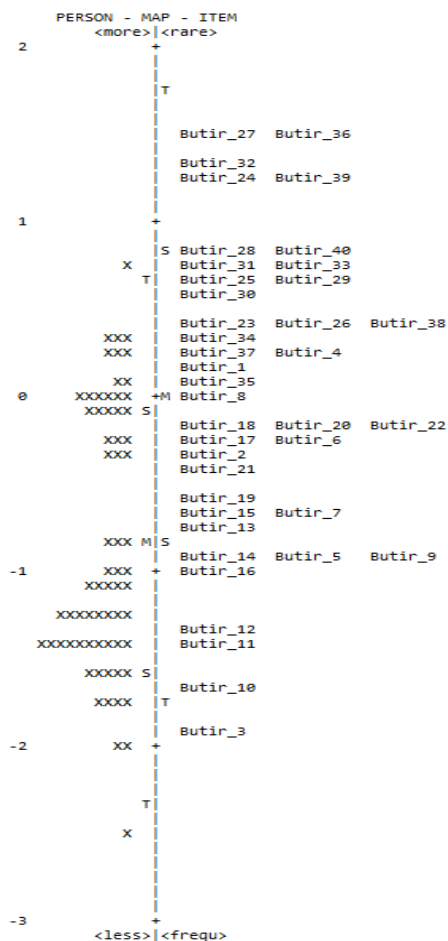


Figure 2. Person Map Item

Table 2. Item Difficulty Level

Item	Item difficulty	Description	Item	Item difficulty	Description
1	.20	Difficult	21	-.40	Easy
2	-.33	Easy	22	-.19	Easy
3	-1.89	Very easy	23	.46	Difficult
4	.28	Difficult	24	1.22	Very difficult
5	-.95	Easy	25	.65	Difficult
6	-.26	Easy	26	.46	Difficult
7	-.68	Easy	27	1.52	Very difficult
8	.04	Difficult	28	.85	Difficult
9	-.88	Easy	29	.65	Difficult
10	-1.66	Very easy	30	.65	Difficult
11	-1.44	Very easy	31	.75	Difficult
12	-1.30	Very easy	32	1.37	Very difficult
13	-.75	Easy	33	.75	Difficult
14	-.95	Easy	34	.37	Difficult
15	-.68	Easy	35	.12	Difficult
16	-1.02	Very easy	36	1.52	Very difficult
17	-.26	Easy	37	.28	Difficult
18	-.19	Easy	38	.46	Difficult
19	-.54	Easy	39	1.22	Very difficult
20	-.19	Easy	40	.85	Difficult

Differential Item Function

In the winsteps program package, information about the bias of this item can be found through Item:DIF, between / within. Items that have a P value (PROB < 0.05) indicate that the item is infected with DIF. In the results of the winsteps analysis, it is known

that the probability value (PROB < 0.05) which means that infection is biased between men and women in five items, namely item 24 (PROB = 0.0326), item 32 (PROB = 0.0351), item 33 (PROB = 0.0262), item 38 (PROB = 0.0028), and item 39 (PROB = 0.0462). Furthermore, it can be seen in Figure 2.

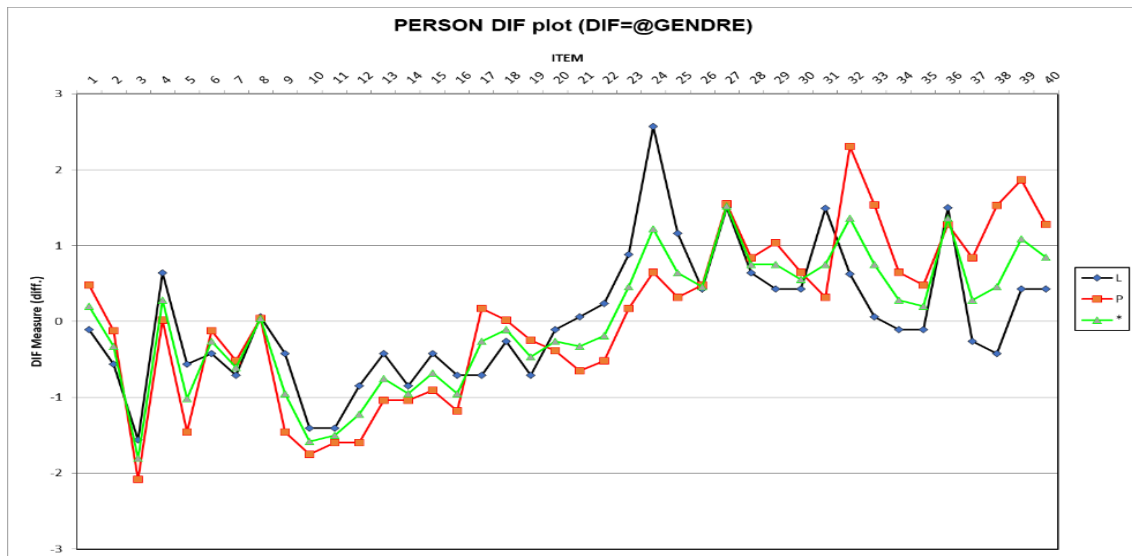


Figure 3. Person DIF Plot

Based on Figure 3, the level of difficulty of the relative items for each group. The higher the graph point, the more difficult the item is for the group. There are three curves based on gender, namely L (male), P (female), and an * (star) which indicates the average value. From Figure 4, it can be seen roughly that the distance of the DIF measure value between L and P is the furthest at points 24, 32, 33, 38, and 39. While in other items the distance between L and P is not too far. This shows that the five items have quite a large difference in difficulty levels between men and women. In item 24, men benefit more because the item is more difficult for women than men. For items 32, 33, 38, and 39 women benefited more because these items were more difficult for men than women. Therefore, the five items should be reviewed whether it is true that these items are more profitable for women and men.

CONCLUSION

The test instrument used for the final exam in mathematics is fit with the Rasch model. This is indicated by an item score (item reliability) of 0.87, person reliability (person reliability) of 0.72, and Cronbach's alpha value of 0.77 while the Outfit Mean Square Statistic (Outfit MNSQ) value of 1.09 in the person and item columns. The Outfit Z Standard (Outfit ZSTD) value is -0.1 in the person table and -0.2 in the item table. While the number of items that fit as many as 35 while those who do not fit as many as 5 items. The researchers suggest further research to calculate the quality of items using other models and approaches, such as 2PL, 3PL or 4PL with different software applications.

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.
- Abdullah, R. (2017). Urgensi Penilaian Hasil Belajar Berbasis Kelas Mata Pelajaran Ips Di Madrasah Tsanawiyah. *Lantanida Journal*, 3(2), 168.
- Agustin, N., Susilogati, S., & Addiani, K. (2018). Desain Instrumen Tes Bermuatan Etnosains Untuk Mengukur Kemampuan Berpikir Kritis Siswa Sma. *Jurnal Inovasi Pendidikan Kimia*, 12(2).
- Alfarisa, F., & Purnama, D. N. (2019). Analisis Butir Soal Ulangan Akhir Semester Mata Pelajaran Ekonomi SMA Menggunakan RASCH Model. 11(2).
- Aprita, Y. M., & Haryati, T. (2021). Analisis Tingkat Kesesuaian Butir Soal (Item Fit) Menggunakan Model Rasch. *Jurnal AKRAB JUARA*, 6(1), 180–193.
- B, M. (2017). Evaluasi Belajar Peserta Didik (Siswa). *Idaarrah: Jurnal Manajemen Pendidikan*, 1(2), 257–267.
- Bambang, S. (2014). *Model Rasch untuk Penelitian Sosial Kuantitatif*.
- Chan, S. W., Looi, C. K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: a Rasch model measurement analysis. *Journal of Computers in Education*, 8(2), 213–236.
- Che Lah, N. H., Tasir, Z., & Jumaat, N. F. (2021). Applying alternative method to evaluate online problem-solving skill inventory (OPSI) using Rasch model analysis. *Educational Studies*, 00(00), 1–23.
- Dwinata, A. (2019). Analisis Kemampuan Pemecahan Masalah Menggunakan Pemodelan RASCH pada Materi Permutasi dan Kombinasi. *PRISMA, Prosiding Seminar Nasional Matematika*, 2, 124–131.
- Falani, I., & Kumala, S. A. (2017). Kestabilan Estimasi Parameter Kemampuan Pada Model Logistik Item Response Theory Ditinjau Dari Panjang Tes. *SAP (Susunan Artikel Pendidikan)*, 2(2), 146–150.
- Fauziana, A., & Wulansari, A. D. (2021). Analisis Kualitas Butir Soal Ulangan Harian di Sekolah Dasar dengan Model Rasch. 6, 10–19.
- 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), 27–32.
- Hambelton, R. K., & Swaminathan, H. (2013). *Item Response Theory: Principles and Applications*. Springer Science & Business Media.
- Imaroh, N., Susongko, P., & Isnani. (2020). Uji Validitas Tes Ulangan Akhir Semester Gasal Mata Pelajaran

- Matematika (Studi Deskriptif Analisis Dokumenter di SMP Negeri Slawi Tahun Pelajaran 2016/2017). *Jurnal Pendidikan MIPA Pancasakti*, 1(1), 80–89.
- Isnani, I., Utami, W. B., Susongko, P., & Lestiani, H. T. (2019). Estimation of college students' ability on real analysis course using Rasch model. *Research and Evaluation in Education*, 5(2), 95–102.
- Liang, T., Wells, C. S., & Hambleton, R. K. (2014). An Assessment of the Nonparametric Approach for Evaluating the Fit of Item Response Models. *Journal Of Educational Measurement*, 51(1), 1–17.
- Mardapi, D. (2012). *Pengukuran Penilaian dan Evaluasi Pendidikan*. Nuha Medika.
- Muharromah, M. M., & Humaisi, S. (2020). Analisis Butir Soal Penilaian Ujian Semester Gasal Mata Pelajaran IPS di MTs Darul Muna Ponorogo. *ASANKA: Journal of Social Science And Education*, 1(2), 102–114.
- Muluki, A., Bundu, P., & Sukmawati, L. (2020). Analisis Kualitas Butir Tes Semester Ganjil Mata Pelajaran IPA Kelas IV Mi Radhiatul Adawiyah. *Jurnal Ilmiah Sekolah Dasar*, 4(1), 86–96.
- Nielsen, T., Martínez-García, I., & Alastor, E. (2021). Critical thinking of psychology students: A within- and cross-cultural study using Rasch models. *Scandinavian Journal of Psychology*, 62(3), 426–435.
- Nufus, S. H., Gani, A., & Suhendrayatna. (2017). Pengembangan Instrumen Penilaian Sikap Berbasis Kurikulum 2013 Pada Pembelajaran Kimia SMA. *Jurnal Pendidikan Sains Indonesia*, 05(01), 44–51.
- Nurjannah, N. (2017). Efektivitas Bentuk Penilaian Formatif Disesuaikan Dengan Media Pembelajaran. *PARAMETER: Jurnal Pendidikan Universitas Negeri Jakarta*, 29(1), 75–90.
- Palinussa, A. L., & Thaib, A. (2020). Mengkaji Perbedaan Kemampuan Matematis Siswa Menggunakan Model Pembelajaran Kooperatif Think Pair Share (TPS) dan Model Pembelajaran Konvensional di Kelas VIII SMP Ditinjau dari Level Sekolah. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 14(1), 23–30.
- Pratama, D. (2020). Analisis Kualitas Tes Buatan Guru Melalui Pendekatan Item Response Theory (IRT) Model Rasch. *Tarbawy: Jurnal Pendidikan Islam*, 7(1), 61–70.
- Retnawati, H. (2014). *Teori respons butir dan penerapannya: Untuk peneliti, praktisi pengukuran dan pengujian, mahasiswa pascasarjana*. Nuha Medika.
- Salman, A., & Abd.aziz, A. (2015). Evaluating user Readiness towards Digital Society: A Rasch Measurement Model Analysis. *Procedia Computer Science*, 65(Iccmit), 1154–1159.
- Santoso, B. (2014). Penilaian Portofolio Dalam Matematika. *Jurnal Pendidikan Matematika*, 1(2), 31–38.
- Scoulas, J. M., Aksu Dunya, B., & De Groote, S. L. (2021). Validating students' library experience survey using rasch model. *Library and Information Science Research*, 43(1), 101071.
- Solichin, M. (2017). Analisis Daya Beda Soal Taraf Kesukaran, Butir Tes, Validitas Butir Tes, Interpretasi Hasil Tes Validitas Ramalan dalam Evaluasi Pendidikan. *Jurnal Manajemen Dan Pendidikan Islam*, 2(2), 192–213.
- Sumaryanto, T., & Khumaedi, M. (2019). the Analysis of Instrument of the Ability To Acting and Thinking Creatively Based Rasch Model. *Journal of Research and Educational Research Evaluation*, 8(1), 48–56.
- Sumintono, B. (2014). Model Rasch untuk Penelitian Sosial Kuantitatif. *ITS Surabaya, November 201*, 1–9.
- Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial*. Trim Komunika.
- 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.
- Sutama, S., Sandy, G. A., & Fuadi, D. (2017). Pengelolaan Penilaian Autentik Kurikulum 2013 Mata Pelajaran Matematika di SMA. *Manajemen Pendidikan*, 12(1), 105.
- Talakua, M., Ratumanan, T. G., & Tamalene, H. (2020). Komparasi Hasil Belajar Siswa Yang Menggunakan Media Swishmax Dan Tanpa Swishmax Pada Materi Operasi Hitung Pecahan Di Kelas Vii Smp Negeri 3 Ambon. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, 14(1), 031–038.
- Tseng, M. C., & Wang, W. C. (2021). The Q-Matrix Anchored Mixture Rasch Model. *Frontiers in Psychology*, 12(March), 1–9.
- Umami, M. (2018). Penilaian Autentik Pembelajaran Pendidikan Agama Islam dan Budi Pekerti dalam Kurikulum 2013. *Jurnal Kependidikan*, 6(2), 222–232.

- van der Linden, W. J., & Hambleton, R. K. (2013). *Handbook of modern item response theory*. Springer Science & Business Media.
- Wei, S., Liu, X., Wang, Z., & Wang, X. (2012). Using Rasch Measurement To Develop a Computer Modeling-Based Instrument To Assess Students' Conceptual Understanding of Matter. *Journal of Chemical Education*, 89(3), 335–545.