Jurnal Elemen, 8(1), 117-130, January 2022
https:/doi.org/10.29408/jel.v8i1.4482

e- ISSN: 2442 - 4226

Assessing the item of final assessment mathematics test of junior high school using Rasch model

Atikah [1]*, Sudiyatno [2], Abdul Rahim [1], Marlina [1]

[1] Department of Educational Research and Evaluation, Yogyakarta State University, Yogyakarta, Indonesia

[2] Department of Mechanical Engineering Education, Yogyakarta State University, Yogyakarta, Indonesia

* Correspondence: atikah0609pasca.2019@student.uny.ac.id

Abstract

The test is a tool to collect information about the achievement of learning objectives so that the test must have good quality in order to be able to measure students' abilities accurately. To determine the quality of the test, it is necessary to do an analysis, one of which is the Rasch model approach. This study aims to analyze the quality of the final semester test results in learning mathematics—sampling using a purposive sampling technique. The subjects of this study were 56 students of Junior High School (SMP) in Yogyakarta, consisting of 25 males and 31 females. The test instrument consists of 40 multiple choice questions with the correct answer given a score of 1 and the wrong answer being given a score of 0. Therefore, the data obtained is dichotomous. The test kits were analyzed by applying the Rasch model. The study results obtained an estimate of the validity of item fit, where the questions analyzed had an INFIT MNSQ value between 0.94 - 1.11 and all questions with an OUTFIT value $t \leq 2.00$; thus, 40 items fit according to the Rasch model. The characteristics of the test questions based on the Rasch model have a very difficult difficulty level of 0 items by 0%. The difficult category is three items at 7,5%, the medium category is 20 items at 50%, the easy category is 17 items at 42.5%, and 0% for the very easy category. The reliability of the case estimate value is 0.00 in the medium category, and the reliability of the item estimate value is 0.85 in the very high category. The implication of this research is additional knowledge regarding the Rasch analysis of mathematical problem models as a graduation standard.

Keywords: final assessment; item response theory; Rasch model

Received: 16 November 2021 | Revised: 9 December 2021
Accepted: 18 December 2021 | Published: 6 January 2022

Introduction

Assessment is a component that is always attached to the teaching and learning process. The assessment results can be used as a reference for teachers to determine success and improve the quality of teaching (Mardapi, 2016). Assessment is also a series of information-gathering activities related to student achievement (Stiggins, & Chappuis, 2012). In addition, assessment is an activity to interpret the measurement results (Santoso et al., 2019). In the assessment, there are three objectives of the assessment process in the learning process, namely, diagnosing students' learning difficulties, measuring improvements over time, and digging up information that students can use to improve their achievements. (Dunn et al., 2003). So a good assessment requires a good test instrument, too (Agustin et al., 2018; Nufus et al., 2017).

The test is a tool to collect information about the achievement of educational goals or learning objectives (Wahyudi, 2012). Besides that, the test is also a specific method that can be used or a procedure that needs to be taken in the context of measurement and assessment in education (Kadir, 2015)**.** The tendency of teachers in compiling learning outcomes tests usually uses existing tests (Nazaruddin, 2017). In addition to the validity and reliability criteria, the instrument in the form of a test needs to meet other criteria, including the level of difficulty, discriminating power, and distracting power (Iskandar & Rizal, 2018; Kones et al., 2021). In contrast, the deceptive power informs about the functioning of the answer options (Amiruddin et al., 2020). *Classical test theory* is an approach that assumes that the observed score is the sum of the pure scores and measurement errors (Himelfarb, 2019). Classical test theory is also the dominant approach used (Petrillo et al., 2015).

Test takers with high abilities have a greater chance of answering correctly than test-takers with low abilities (Retnawati, 2014; Hambleton, R. K. Swaminathan, H., & Rogers, 1985; Hambleton et al., 1991). Three assumptions form IRT's basis: unidimensionality, local independence, and parameter invariance (Retnawati, 2014). In this case, the response or performance of the test taker is something that can be observed (observable). At the same time, the nature of ability is not visible (unobservable), which underlies the performance on the test. (Embretson, E. & Reise, 2000). Item response theory as an independent test. The purpose of developing item response theory is to overcome the weakness of classical test theory, which is not independent of the group of participants who take the test and the test being tested (Embretson, E. & Reise, 2000). The relationship between success in each item with a person's ability is described by an increasing monotonic function called the item characteristic function (Lord, 1980).

The Rasch model is a modern valuation theory that can classify item and person calculations in a distribution map (Halim et al., 2001). The Rasch model is a modern valuation theory that can classify item and person calculations in a distribution map (Thissen et al., 2001). The stringent requirements that must be met have caused many researchers to find that the items analyzed do not fit the Rasch model. (Goldstein, 1982). Two strategies are commonly used when the researcher finds that the data is not sufficiently supportive to be approached using the Rasch model (Smit et al., 2000). In the Rasch model approach, besides paying attention to items,

they also pay attention to aspects of response and correlation (Ardiyanti, 2016). There are several advantages of IRT, namely; 1) the score describes the ability of the test taker and does not depend on the difficulty of the test, 2) it can be used to connect item items with the ability of test-takers, 3) does not require parallel tests to determine the reliability coefficient (Andayani et al., 2019). The Rash model also calibrates three things: the measurement scale, the test participant (person), and the item. Calibration is intended to ensure the validity and reliability of the measurement results so that the test can provide comprehensive information. (Ardiyanti, 2016). Item analysis with Rasch modeling will produce information about the characteristics of items and students that have been formed into the same metric (Sumintono & Widhiarso, 2015).

Analysis of test instruments using the Rash model can be done through the following steps; 1) assess item fit statistic. This stage is the stage to determine the items that match the Rasch model. If items do not match, the Rash model's analysis of test instruments can be done through the following steps; 1) assess item fit statistic. (2) match the Rasch model, and 3) Determine which items and test-takers (person) fit the Rasch model through the goodness of fit analysis (Mardapi, 2016). One of the simplest models and has been widely used by experts in developing a test is the Rasch model, with one parameter (1PL). In addition, the selection of the Rasch model is because this model has at least fulfilled the principles of the measurement model, namely; 1) this model can provide linear measures with equal intervals, 2) can overcome the problem of missing data, 3) can provide more precise estimates, 4) can detect the imprecision of a model, and 5) provide measurement instruments that are independent of parameters researched (Mardapi, 2016). Measurements in the Rasch model are direct comparisons between individuals and items. The Rash model applying test instrument analysis has advantages over other models such as classical theory (Sumintono & Widhiarso, 2015).

In education, test quality analysis has been carried out a lot, so many studies have replaced the classical theory approach with the item response theory (IRT) approach. The following are some studies that use the Rasch model in mathematics learning research conducted by (Alfarisa & Purnama, 2019; Rahim & Haryanto, 2021) explained that the Rasch model analysis could provide information about the level of difficulty of the questions. Even in research of Imaroh et al. (2020), analyzing the items using the Rasch model can provide information about the quality of the items on the odd semester final test of mathematics for class VII Junior High School (SMP). While in research, Azizah and Wahyuningsih (2020) analyze the test instrument used to measure students' abilities in actuarial mathematics courses at the Mathematics Department, State University of Malang. Rasch model is also used in research (Erfan et al., 2020).

Based on previous research in this study, this study aims to develop a valid and reliable test that can be used in analyzing the quality of the final semester exam test in mathematics learning. In addition, to determine the feasibility of the test instrument and describe the characteristics of the test items for the math semester final exam using the Rasch model.

Methods

This study uses quantitative research methods exploratory, descriptive approach. The selection of this method is based on data taken from the responses of students' answers to the 8th-grade junior high school mathematics exam, which then researchers describe or describe the data collected as they are with research data in the form of numbers. The exploratory description approach is part of the descriptive research method (Saebani, 2015). Sampling using a purposive sampling technique. The subjects of this study were 56 students of Junior High School (SMP) in Yogyakarta, consisting of 25 males and 31 females. The research object is in the form of test equipment and student responses in the final semester exam for mathematics. The data collection technique used in this study is a test that is packaged in a google form. The indicators on the test instrument consist of Number Patterns, Cartesian Coordinates, Relations and Functions, Straight Line Equations, Two-Variable Linear Equations Systems. The test instrument consists of 40 multiple choice questions with the correct answer given a score of 1 and the wrong answer being given a score of 0.

The test device was analyzed quantitatively based on the Item Response Theory (IRT) approach applying the Rasch model. The Rasch model is used to analyze the response of the items and the relationship between the students' ability level and the difficulty level of the item items. The instrument is designed from variables that have been defined satisfactorily. The relevant constructs are identified, from which the item items are created and developed to measure the desired variable (Sumintono & Widhiarso, 2015). The validity of using the Rasch model, which can provide statistics, is also to investigate the validity of the test instrument based on the response of the research subject. Analysis of the questions with the help of the QUEST program was then seen by the fit model, index of difficulty level, and reliability. Based on the item analysis results, the quality of the items is empirically categorized as good or bad by applying the Rasch model.

Results

Estimated Item Validity

The criteria for the INFIT MNSQ value can be seen in table 1. The results of the INFIT MNSQ value analysis in the QUEST program recapitulation of item validity can be seen in Table 2. The results show information in item validity for all items classified as fit or matching the Rasch model, which ranges from INFIT values MNSQ 1.

The validity of the test instrument was tested using the QUEST program, which, as disclosed, not all test takers can answer correctly on one item of the question. Based on Rasch modeling, if there are items answered correctly or incorrectly by students who take the test, they will not be counted. Therefore, to find out which items match the Rasch Model, you can also find out through the item fit map as shown in Figure 1.

Table 1. INFIT MNSQ score criteria

Mark INFIT MNSQ	Description
>1.33	Does not match the model
0.77-1.33	Fits the model
<0.77	Does not match the model

Tabel 2. Item validity recapitulation

Item	INFIT MNSQ Value	Description	Item	INFIT MNSQ Value	Description
1	0.99	Fits the model	21	0.97	Fits the model
2	0.97	Fits the model	22	0.94	Fits the model
3	1.08	Fits the model	23	1.10	Fits the model
4	1.01	Fits the model	24	0.97	Fits the model
5	0.99	Fits the model	25	0.98	Fits the model
6	1.00	Fits the model	26	0.98	Fits the model
7	1.03	Fits the model	27	0.96	Fits the model
8	1.02	Fits the model	28	0.95	Fits the model
9	0.98	Fits the model	29	0.99	Fits the model
10	1.02	Fits the model	30	1.02	Fits the model
11	1.03	Fits the model	31	0.97	Fits the model
12	1.12	Fits the model	32	1.06	Fits the model
13	0.99	Fits the model	33	0.96	Fits the model
14	0.99	Fits the model	34	0.97	Fits the model
15	1.11	Fits the model	35	1.00	Fits the model
16	0.98	Fits the model	36	0.97	Fits the model
17	0.97	Fits the model	37	0.97	Fits the model
18	1.03	Fits the model	38	0.98	Fits the model
19	1.00	Fits the model	39	0.94	Fits the model
20	1.00	Fits the model	40	0.96	Fits the model

Figure 1. below is the result of the fit map model from the image above. It can be seen that all items are in the range of values in the INFIT MNSQ 0.77 – 1.30. The dots on the left show a value of 0.77, while the dots on the right show a value of 1.30.

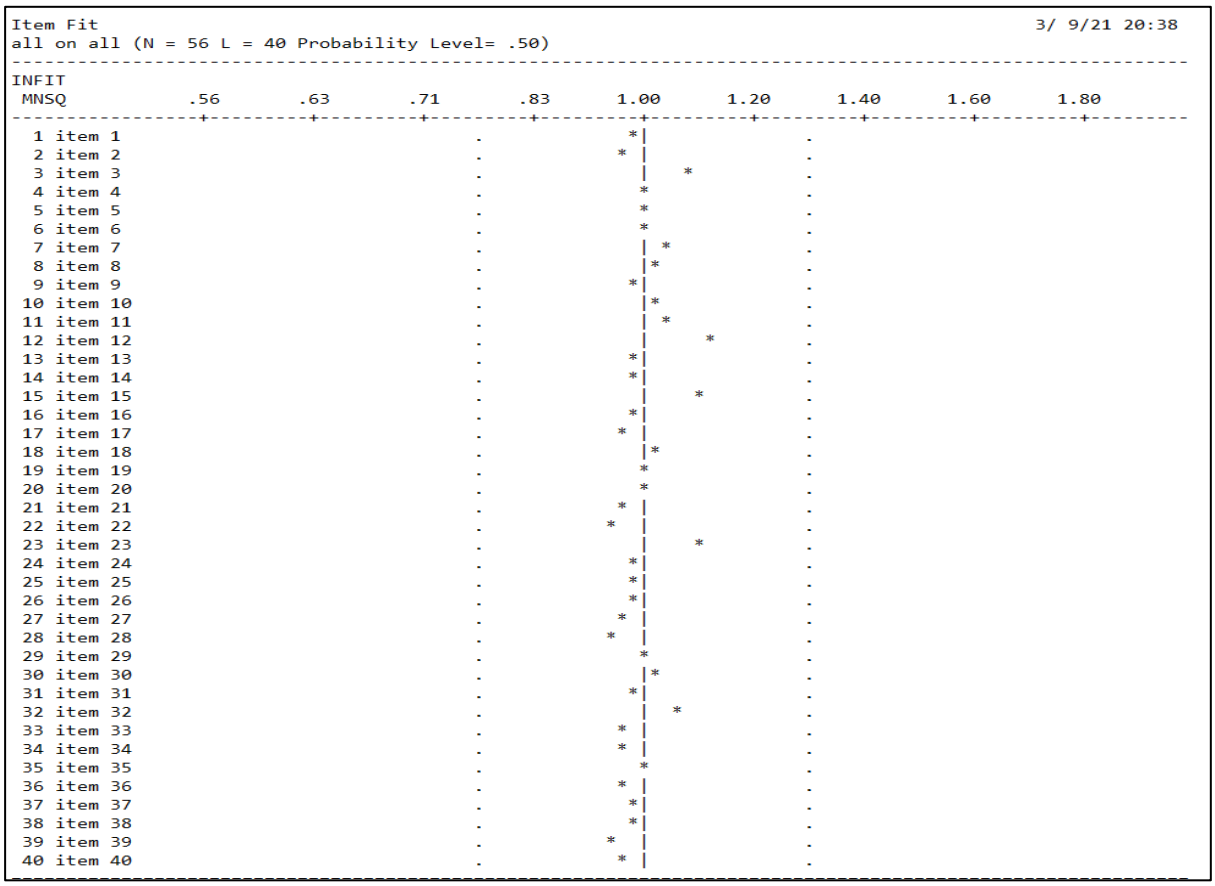


Figure 1. Fit map of Rasch model

Difficulty Estimation

The item difficulty level through the QUEST program can be obtained by looking at the item estimate (Threshold) analysis results. The criteria in determining the difficulty level of items on a test with a value range of -2.0 – 2.0. If the distribution range of the test or test takers is <-2.0, then the items on the test are included in the easy category. If the range of items or test takers is >2.0, then the items on the test are in the difficult category. To find out more accurately, the range value and the distribution of the item difficulty level can be seen in Figure 2.

Figure 2 shows the distribution of the level of difficulty of the item, it can be obtained that item number 9 is the item that is categorized as the most difficult. When compared with the ability of students taking the test, the possibility of students taking the test correctly answering item number 9 is very small, it can be said to be impossible. While item number 23 is the item with the easiest category and according to the students' abilities taking the test. The difficulty level of items through the QUEST program can be seen from the item estimate threshold with the item estimate threshold value (Setyawarno, 2016).

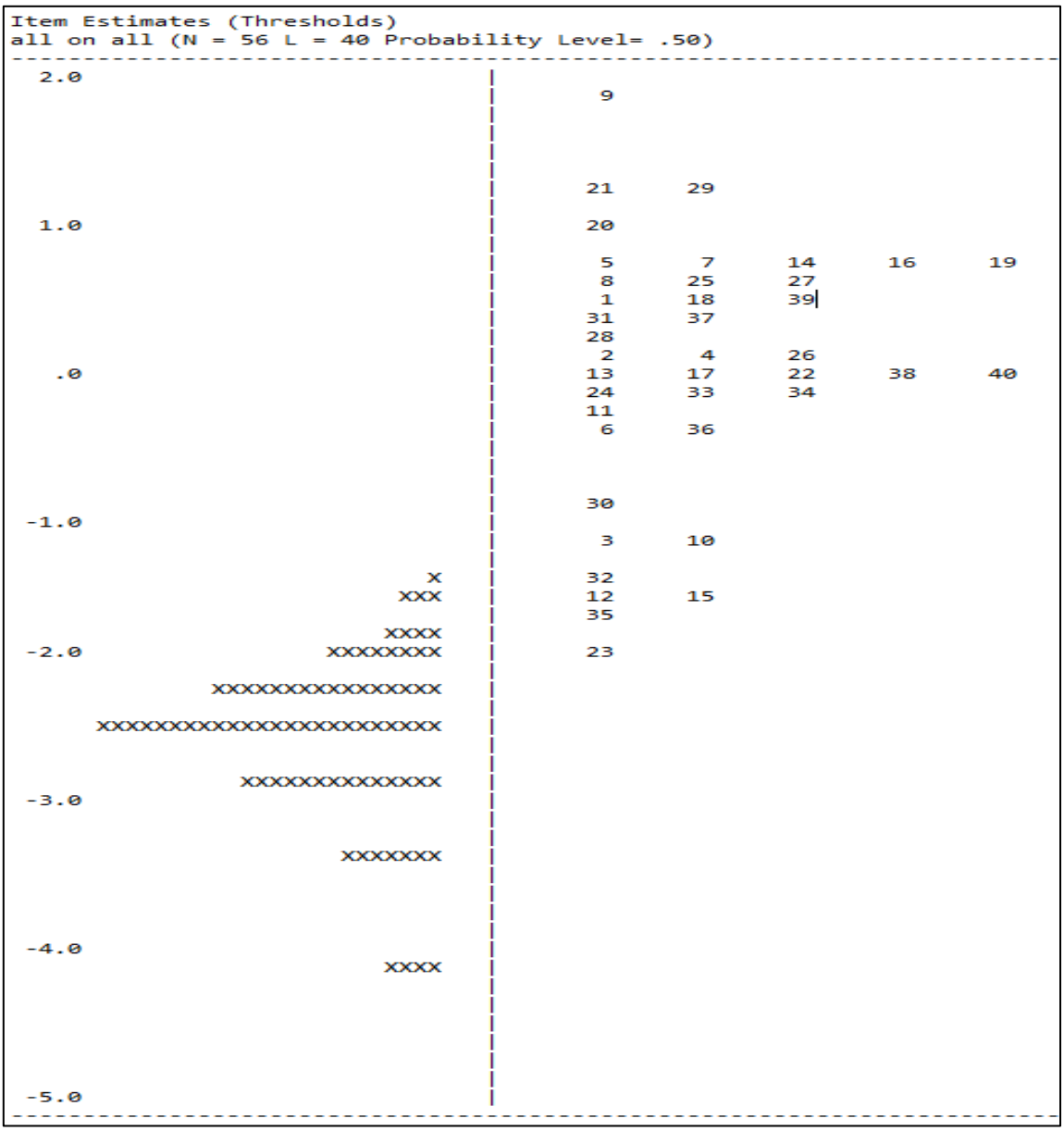


Figure 2. Item difficulty level distribution

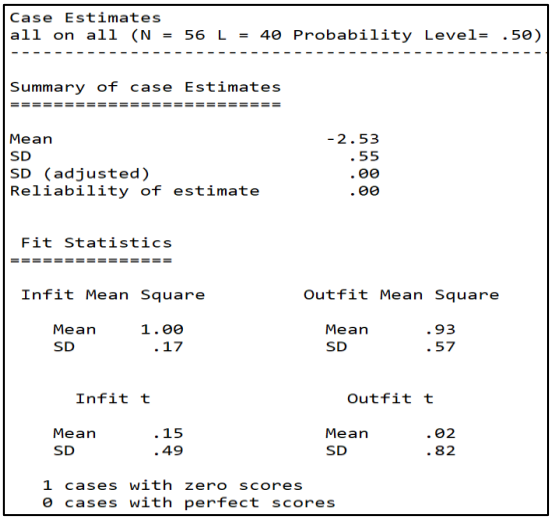
Table 3. Threshold value item estimate

Threshold Value	Criteria
b>2	Very difficult
1<b≤2	Difficult
-1≤b≤1	Currently
-2 ≤b<-1	Easy
b<-2	Very easy

Table 4. Recapitulation of the difficulty level of the Rasch model questions

Item	Threshold Value	Interpretation	Item	Threshold Value	Interpretation
1	0.47	Medium	21	1.18	Difficult
2	0.04	Medium	22	-0.04	Easy
3	-1.24	Easy	23	-1.88	Easy
4	0.04	Medium	24	-0.20	Easy
5	0.77	Medium	25	0.61	Medium
6	-0.40	Easy	26	0.14	Medium
7	0.77	Medium	27	0.61	Medium
8	0.61	Medium	28	0.24	Medium
9	1.87	Difficult	29	1.18	Difficult
10	-1.13	Easy	30	-0.88	Easy
11	-0.34	Easy	31	0.35	Medium
12	-1.61	Easy	32	-1.49	Easy
13	-0.04	Easy	33	-0.13	Easy
14	0.77	Medium	34	-0.13	Easy
15	-1.61	Easy	35	-1.68	Easy
16	0.77	Medium	36	-0.47	Medium
17	-0.04	Easy	37	0.35	Medium
18	0.47	Medium	38	-0.04	Medium
19	0.77	Medium	39	0.47	Medium
20	0.95	Medium	40	-0.04	Easy

Based on table 4, recapitulation of the difficulty level of each item. The results of the difficulty level can be seen that for items in the very difficult category, as many as 0 items are 0%. The difficult category has three items at 7,5%, the medium item category is 20 items at 50%, the easy category is 17 items at 42,5%, and 0% for the very easy item category. The overall ability of test-takers is below the item difficulty level. It can be proven by the small number of test-takers who correctly answered very difficult or difficult items. How to find out the ability of test-takers through the QUEST program, the analysis value is seen in the Summary of Case Estimate on the reliability of estimate with criteria, if the Estimate value is > 1.00 in the high ability category, -1.00 – 1.00 in moderate ability, and < -1.00 in low ability.



```
Case Estimates
all on all (N = 56 L = 40 Probability Level= .50)
-------------------------------------------------

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

Mean                          -2.53
SD                              .55
SD (adjusted)                   .00
Reliability of estimate         .00


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

 Infit Mean Square          Outfit Mean Square

    Mean    1.00               Mean      .93
    SD       .17               SD        .57


      Infit t                    Outfit t

    Mean     .15               Mean      .02
    SD       .49               SD        .82

   1 cases with zero scores
   0 cases with perfect scores
```

Figure 3. Estimation of respondent's ability

Based on Figure 3, it can be obtained information that the students on the test have the moderate ability, with a reliability estimate value of 0.00. To find out the reliability of the item estimate is presented in Figure 4.

```
Item Estimates (Thresholds)
all on all (N = 56 L = 40 Probability Level= .50)
-------------------------------------------------

Summary of item Estimates
=========================

Mean                              .00
SD                                .88
SD (adjusted)                     .81
Reliability of estimate           .85


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

 Infit Mean Square          Outfit Mean Square

    Mean     1.00              Mean      .93
    SD        .04              SD        .21


      Infit t                    Outfit t

    Mean      .21              Mean     -.05
    SD        .46              SD        .65

   0 items with zero scores
   0 items with perfect scores
```

Figure 4. Reliability of item estimate

Based on Figure 4, the reliability of the item estimate is obtained at 0.85 with a very high category. The reliability value based on item estimates is also called sample reliability. The higher the value, the more items that fit or match the tested model. Conversely, the lower the value, the more items that do not fit or match the model being tested, so it does not provide the expected information.

Estimated Items Passed (Fit)

To find out which items fall or pass, it is based on the OUTFIT t value in the QUEST program. If the value of OUTFIT t $\leq$ 2.00, then the item passes, and if the value of OUTFIT t $\geq$ 2.00, the item fails. Based on table 5, the results where all items pass, it can be concluded that all items can be used. In addition, it is better not to be included in the test for items with the highest difficulty level and the easiest. As for the test takers' ability, few of the test takers can answer correctly on the most difficult item. Then the proportion of difficult items is reduced to keep up with the test takers. Based on the analysis results, the test takers belong to the category with moderate ability.

Table 5. Recapitulation of fit items

Item	Outfit t Value	Description	Item	Outfit t Value	Description
1	-0.4	Fit	21	-0.7	Fit
2	-0.6	Fit	22	-0.6	Fit
3	0.6	Fit	23	1.5	Fit
4	-0.1	Fit	24	-0.5	Fit
5	-0.2	Fit	25	0.1	Fit
6	0.0	Fit	26	-0.3	Fit

Item	Outfit t Value	Description	Item	Outfit t Value	Description
7	1.0	Fit	27	-0.7	Fit
8	0.3	Fit	28	-0.7	Fit
9	-0.4	Fit	29	-0.1	Fit
10	-0.1	Fit	30	0.0	Fit
11	0.6	Fit	31	-0.4	Fit
12	1.6	Fit	32	0.4	Fit
13	-0.1	Fit	33	-0.7	Fit
14	0.0	Fit	34	0.2	Fit
15	1.1	Fit	35	-0.3	Fit
16	0.0	Fit	36	-0.8	Fit
17	-0.3	Fit	37	-0.7	Fit
18	1.5	Fit	38	-0.2	Fit
19	-0.2	Fit	39	-1.0	Fit
20	0.0	Fit	40	-0.7	Fit

Discussion

Item analysis helps improve the quality of items through revisions or discarding ineffective questions. Besides that, it can be used as diagnostic information for students, whether they have understood the material that has been taught (Fauziana & Wulansari, 2021). Analysis of the 40 items using the item response theory Rasch model approach obtained all items fit with the Rasch model with an INFIT MNSQ value between 0.94 - 1.11 and an OUTFIT t value obtained ≤ 2.00; thus, 40 items matched and passed accordingly with rash models. According to research conducted by (Widyaningsih et al., 2021), the items that match the Rasch model have INFIT MNSQ values in the range 0.77 - 1.33 and OUTFIT values t ≤ 2.00.

In addition to the fit model with the Rasch model, there are also grain characteristics. Based on the results of the analysis of the characteristics of the test items based on the Rasch model, the level of difficulty in the very difficult category is 0 items by 0%. Items with a difficult category as much as three items are 7,5%, the medium item category is 20 items at 50%, the easy category is 17 items at 42,5%, and 0% for the very easy item category. Analysis of the reliability value with the Rasch model using the QUEST program to determine the reliability of item estimate and case estimate. The reliability of the item estimate is 0.85. Reliability with the Rasch model 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) are as follows; <0.67 weak, 0.67-0.80 moderate, 0.81 – 0.90 good, 0.91 – 0.94 very good, >0.94 perfect. The reliability of the item estimate is 0.85. Relates to the number of items that fit the Rasch model. The value of 0.85 is classified as reliability with a very good category so that it affects the items that fit the Rasch model. So the higher the reliability, the more items fit with the Rasch model. Reliability of estimate to determine student ability with criteria, if Estimate value > 1.00 in high ability category, -1.00 – 1.00 moderate ability, and < -1.00 low ability (Pratama, 2020). The reliability of the case estimate value of 0.00 belongs to the medium category. A value that indicates an inconsistency as expressed (Ardiyanti, 2016) on the test taker's answers. The inconsistency of the answers of the test takers can also mean that the test

takers are careless in answering the questions, affecting the reliability value of the person/subject to be low. The value of the reliability of the case estimate is in a good category; the answers to the test-takers show their consistency. If the value of the reliability of the case estimate is low which is influenced by the number of test participants. The answers of the test takers are less than 100, which is 56.

The inconsistency of the test takers' answers can also mean that the test takers are careless in answering the questions, affecting the reliability value of the person/subject to be low. The value of the reliability of the case estimate is in a good category; the answers to the test-takers show their consistency. If the value of the reliability of the case estimate is low, which is influenced by the number of test participants. The test takers' answers are less than 100, which is 56. In this case, the number of test-takers in the study of Purba (2018) as many as 428 students with a reliability value of 0.92. The study of Hakiki et al. (2018) also shows the same result that the number of items does not affect the test reliability value. The number of items in this study was 20 items, with 293 respondents. It is in line with the research conducted by Istiyono et al. (2014), where the number of items analyzed using the Rasch model is 26 items, and the number of test-takers is 1001 students. Research conducted (Purba, 2018) Regarding the analysis of the achievement test instrument using the Rasch model, it is proven that the number of test participants < 100 affects the reliability value of the test takers. In addition, based on the research findings, it can also be understood that the number of item items does not affect the test taker's reliability score.

Conclusion

The quality of the final exam of the mathematics test using the Rasch model obtained an estimate of the validity of the item fit. The 40 questions analyzed had an INFIT MNSQ value between 0.94 - 1.11, and all questions with a score of OUTFIT t were obtained ≤ 2.00. Thus 40 items matched and passed according to the rash model. The characteristics of the test questions, in general, are pretty good. The characteristics of quantitative test questions based on the item response theory approach based on the Rasch model have a very difficult category difficulty level of 0 items by 0%. Items with three items in the difficult category are 7,5%, the medium item category is 20 items at 50%, the easy category is 17 items at 42,5%, and 0% for the very easy item category. The reliability estimate value is 0.00 with a medium category, and the reliability of the item estimate value is obtained at 0.85 with a very high category.

Conflicts of Interest

The authors declare that no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely by the authors.

References

Agustin, N., Sudarmin, Sumarti, S. S., & Addiani, A, K. (2018). Desain instrumen tes bermuatan etnosains untuk mengukur kemampuan berpikir kritis siswa SMA [Design of an ethnoscience-laden test instrument to measure high school students' critical thinking skills]. *Jurnal Inovasi Pendidikan Kimia*, *12*(2), 2159-2169. https://journal.unnes.ac.id/nju/index.php/JIPK/article/view/15475

Alfarisa, F., & Purnama, D. N. (2019). Analisis butir soal ulangan akhir semester mata pelajaran ekonomi SMA menggunakan RASCH model [Analysis of the final exam items for high school economics subjects using the RASCH model]. *Jurnal Pendidikan Ekonomi*, *11*(2), 366-374. https://ejournal.undiksha.ac.id/index.php/JJPE/article/view/20878

Amiruddin, K., Mania, S., Ichiana, N. N., Majid, A. F., Tarbiyah, F., Islam, U., & Alauddin, N. (2020). Analisis butir soal ujian akhir sekolah (UAS) mata pelajaran matematika [Analysis of items for the final school exam (UAS) for mathematics subjects]. *Alauddin Journal of Mathematics Education*, *2*(2), 207–217. http://journal.uin-alauddin.ac.id/index.php/ajme/article/view/17364

Andayani, A., Purwanto, & Ramalis, T. R. (2019). Kajian implementasi teori respon butir dalam menganalisis instrumen tes materi fisika. *Seminar Nasional Fisika*, *1*(1), 37–42.

Ardiyanti, D. (2016). Aplikasi model Rasch pada pengembangan skala efikasi diri dalam pengambilan keputusan karir siswa [Application of the Rasch model on the development of self-efficiency scale in student career decision making]. *Jurnal Psikologi*, *43*(3), 248–263. https://doi.org/10.22146/jpsi.17801

Azizah, & Wahyuningsih, S. (2020). Penggunaan model Rasch untuk analisis instrumen tes pada mata kuliah matematika aktuaria [The use of the Rasch model for analysis of test instruments in actuarial mathematics courses]. *JUPITEK: Jurnal Pendidikan Matematika*, *3*(1), 45–50. https://doi.org/10.30598/jupitekvol3iss1pp45-50

Dunn, L., Morgan, C., O'Reilly, M., & Parry, S. (2003). *The student assessment handbook: New directions in traditional and online assessment*. Routledge. https://doi.org/10.4324/9780203416518

Embretson, E. & Reise, S. . (2000). *Item response theory for psychologists* (L. E. Associates (ed.)). NJ Publications.

Erfan, M., Maulyda, M. A., Hidayati, V. R., Astria, F. P., & Ratu, T. (2020). Tes klasik dan model Rasch [Classic tests and Rasch models]. *Indonesian Journal of Educational Research and Review*, *3*(1), 11–19. https://doi.org/10.23887/ijerr.v3i1.24080

Fauziana, A., & Wulansari, A. D. (2021). Analisis kualitas butir soal ulangan harian di sekolah dasar dengan model Rasch [Analysis of the quality of daily test items in elementary schools using the Rasch model]. *Jurnal Kependidikan Dasar Islam Berbasis Sains*, *6*(1), 10–19.

Goldstein, H. & S. B. (1982). The Rasch model still does not fit. *British Educational Research Journal*, *82*, 167-170. https://doi.org/10.1080/0141192820080207

Hakiki, A. W., Fitri, A. R., & Agung, I. M. (2018). Analisis properti psikometri subtes Merkaufgaben (ME) dengan Rasch model [Analysis of the psychometric properties of the Merkaufgaben (ME) subtest using the Rasch model]. *Jurnal Psikologi*, *14*(1), 40. https://doi.org/10.24014/jp.v14i1.4900

Halim, R. A., Zaharim, A., Rashid, R. a., & Masodi, M. S. (2001). Application of logistic regression model in Rasch measurement to establish a performance index: A case in Audits on Malaysian Institute of higher learning. *The 12th WSEAS Int Conf. on Applied Mathematics, Cairo, Eygpt, 29-31 December*, *March 2014*.

Hambleton, R. K. Swaminathan, H., & Rogers, H. J. (1985). *Item response theory: Principles and application*. Kluwer Inc. https://doi.org/10.1007/978-94-017-1988-9

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol.2). Sage publication.

Himelfarb, I. (2019). A primer on standardized testing: History, measurement, classical test theory, item response theory, and equating. *Journal of Chiropractic Education*, *33*(2), 151–163. https://doi.org/10.7899/JCE-18-22

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) [Test the validity of the odd semester final test for mathematics subjects (descriptive study of documentary analysis at SMP Negeri Slawi for the 2016/2017 academic year)]. *Jurnal Pendidikan MIPA Pancasakti*, *1*(1), 80–89.

Iskandar, A., & Rizal, M. (2018). Analisis kualitas soal di perguruan tinggi berbasis aplikasi TAP [Analysis of the quality of questions in higher education based on the TAP application]. *Jurnal Penelitian dan Evaluasi Pendidikan*, *22*(1), 12–23. https://doi.org/10.21831/pep.v22i1.15609

Kadir, A. (2015). Menyusun dan menganalisis tes hasil belajar [Develop and analyze learning outcomes tests]. *Al- Ta'dib: Jurnal Kajian Ilmu Kependidikan*, *8*(2), 70–81.

Kones, N. I., & Rosnawati, R. (2021). Kualitas butir dan estimasi kemampuan matematika siswa SMP pada soal ujian sekolah [Item quality and estimation of junior high school students' mathematical abilities on school exam questions]. *Jurnal Elemen*, *7*(2), 280–294. https://doi.org/10.29408/jel.v7i2.3054

Lord, M. L. (1980). *Application of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Publisher.

Mardapi, D. (2016). *Pengukuran, penilaian dan evaluasi pendidikan (Edisi 2)* [Educational measurement, assessment and evaluation (Issue 2)]. Parama.

Nazaruddin. (2017). Kemampuan guru dalam menyusun tes hasil belajar melalui workshop di SD Negeri Lamteubee [The ability of teachers in preparing learning outcomes tests through workshops at SD Negeri Lamteubee]. *Serambi Akademika*, *5*(1), 32–42.

Nufus, S. H., Gani, A., & Suhendrayatna. (2017). Pengembangan instrumen penilaian sikap berbasis kurikulum 2013 pada pembelajaran kimia SMA [Development of an attitude assessment instrument based on the 2013 curriculum in high school chemistry learning]. *Jurnal Pendidikan Sains Indonesia*, *5*(1), 44–51.

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health*, *8*(1), 95–104. https://doi.org/10.1016/j.jval.2014.10.005

Pratama, D. (2020). Analisis kualitas tes buatan guru melalui pendekatan item response theory (IRT) model Rasch [Analysis of the quality of teacher-made tests through the Rasch model of item response theory (IRT) approach]. *Tarbawy: Jurnal Pendidikan Islam*, *7*(1), 61–70. https://doi.org/10.32923/tarbawy.v7i1.1187

Purba, S. E. D. (2018). Analisis model Rasch instrumen tes prestasi pada mata pelajaran dasar dan pengukuran listrik [Rasch model analysis of achievement test instruments on basic subjects and electrical measurements]. *Wiyata Dharma: Jurnal Penelitian dan Evaluasi Pendidikan*, *6*(2), 142. https://doi.org/10.30738/wd.v6i2.3393

Rahim, A., & Haryanto. (2021). Implementation of item response theory (IRT) Rasch model in quality analysis of final exam tests in mathematics. *Journal of Educational Research and Evaluation*, *10*(2), 57–65.

Retnawati, H. (2014). *Respons butir dan penerapannya* [*Item response and its application*]. Nuha Medika.

Saebani, B. A. (2015). *Filsafat Ilmu dan metode penelitian* [*Philosophy of Science and research methods*]. Pustaka Setia.

Santoso, A., Kartianom, K., & Kassymova, G. K. (2019). Kualitas butir bank soal statistika (Studi kasus: Instrumen ujian akhir mata kuliah statistika Universitas Terbuka) [Quality of statistical question bank items (Case study: Instrument of the final exam for statistics course at the Universitas Terbuka). *Jurnal Riset Pendidikan Matematika*, *6*(2), 165–176. https://doi.org/10.21831/jrpm.v6i2.28900

Setyawarno, D. (2017). *Upaya peningkatan kualitas butir soal dengan analisis aplikasi Quest* [Efforts to improve the quality of items with Quest application analysis]. Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Negeri.

Setyawarno, D. (2016). *Analisis data pengukuran menggunakan program Quest* [*Analysis of measurement data using the Quest* program]. http://staff.uny.ac.id/sites/default/files/pendidikan/didik-setyawarno-spdsi-mpd/analisis-butir-soal-panduan-singkat-penggunaan-quest.pdf

Smit, A., Kelderman, H., & Van Der Flier, H. (2000). The mixed Birnbaum model: Estimation using collateral information. *MPR-Online*, *5*(January 2000), 31–43.

Stiggins, R., & Chappuis, J. (2012). *Introduction to student invoved assessment for learning (6th ed.)*. Addison Wesley.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan* [*Rasch modeling application in educational assessment*]. Trim Komunikata.

Susdelina, Perdana, S. A., & Febrian. (2018). Analisis kualitas instrumen pengukuran pemahaman konsep persamaan kuadrat melalui teori tes klasik dan Rasch model [Analysis of the quality of measuring instruments for understanding the concept of quadratic equations through classical test theory and the Rasch model]. *Jurnal Kiprah*, *6*(1), 41–48. https://doi.org/10.31629/kiprah.v6i1.574

Thissen, D., Nelson, L., Rosa, K., et al. (2001). *Item response theory for items scored in more than two categories*. Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410604729-9

Wahyudi, W. (2012). Assesment pembelajaran berbasis portofolio di sekolah [Portfolio-based learning assessment in schools]. *Jurnal Visi Ilmu Pendidikan*, *2*(1), 288–297. https://doi.org/10.26418/jvip.v2i1.370

Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2021). The development of the HOTS test of physics based on modern test theory; question modeling through e-learning of moodle LMS. *International Journal of Instruction*, *14*(4), 51–68. https://doi.org/10.29333/iji.2021.1444a