P-ISSN 2527-5615
E-ISSN 2527-5607

KALAMATIKA Jurnal Pendidikan Matematika
Volume 5, No. 1, April 2020, pages 51-60

AN INTRODUCTION TO THE RASCH MEASUREMENT MODEL: A CASE OF MATHEMATICS COMPREHENSIVE TEST

Elizar Elizar[1], Cut Khairunnisak[2]

[1]Universitas Syiah Kuala, Darussalam, Banda Aceh, Indonesia.
elizar@unsyiah.ac.id
[2]Universitas Syiah Kuala, Darussalam, Banda Aceh, Indonesia.
cut.khairunnisak@unsyiah.ac.id

ABSTRACT

Mathematics assessments should be designed for all students, regardless of their background or gender. Rasch analysis, developed based on Item Response Theory (IRT), is one of the primary tools to analyse the inclusiveness of mathematics assessment. However, the mathematics test development has been dominated by Classical Test Theory (CTT). This study is a preliminary study to evaluate the mathematics comprehensive test. This study aims to demonstrate the use of Rasch analysis by assessing the appropriateness of the mathematics comprehensive test to measure students' mathematical understanding. Data were collected from one cycle of mathematics comprehensive test involving 48 undergraduate students of mathematics education department. Rasch analysis was conducted using ACER Conquest 4 software to assess the item difficulty and differential item functioning (DIF). The findings show that the item related to geometry is the easiest question for students, while item concerning calculus as the hardest question. The test is viable to measure students' mathematical understanding as it shows no evidence of Differential Item Functioning (DIF). Gender has been drawn for each of the test items. The assessment showed that the test was inclusive. More application of Rasch analysis should be conducted to create a thorough and robust mathematics assessment.

ARTICLE INFORMATION

Keywords

Rasch model
Item Response Theory
Differential Item Functioning
Item difficulty analysis
Measurement

Article History

Submitted *Feb 15, 2020*
Revised *Apr 12, 2020*
Accepted *Apr 15, 2020*

Corresponding Author

Elizar Elizar
Universitas Syiah Kuala
Darussalam, Banda Aceh, Indonesia
Email: elizar@unsyiah.ac.id

How to Cite

Elizar, E. & Khairunnisak, C. (2020). An Introduction to the Rasch Measurement Model: A Case of Mathematics Education Students Comprehensive Test. *Kalamatika: Jurnal Pendidikan Matematika*, *5*(1), 51-60.

https://doi.org/10.22236/KALAMATIKA.vol5no1.2020pp51-60

INTRODUCTION

Assessment is an integral part of teaching and learning and therefore plays a vital role in education system for each educational level around the world. Evaluation of the teaching and learning outcomes may include assessment of learning, assessment as learning, and assessment for learning. Assessment of learning, usually known as a summative test, aims to describe how well the students perform in a particular subject and to provide evidence of their achievement for the students themselves, parents and other stakeholders, including the government (Gardner, 2012). Assessment for learning, also called formative assessment, provides teachers opportunities to investigate the students' current proficiency, including their strengths, weaknesses, and misconceptions. Formative assessment enables teachers to provide feedback for students and modify the learning method, strategies, and resources to improve students learning experience which in turn will enhance students' achievement (Black et al., 2003). Unlike the two previous types of assessment (assessment of learning and assessment as learning) in which teachers play a major role, assessment as learning mainly involves students to self-evaluate their learning and articulate their thinking (Boud & Falchikov, 2007). Assessment, in a broader sense means evaluation, that is closely related to testing items. It is paramount to ensure that the test items measure the trait intended to measure. The most common practice of assessment in many countries, including Indonesia, is using raw data instead of measures. This practice is derived from the Classical Test Theory, where the person statistics depends on the items, and the item statistics such as item difficulty and item discrimination rely on the examinee (Guler et al., 2014). For example, a researcher or teacher conducts a test or survey, reports how many respondents respond or provide the correct solution, and investigate the relationships between each item as well as the correlation of each item and the total score (Bond & Fox, 2016). This does not mean that the statistical analysis described is inappropriate, but the highlight is that examining the relationship between variables is no longer sufficient in developing an instrument. More attention concerning the nature of the data or measures are necessary to successfully create quality scientific items or measures (Bond & Fox, 2016). In brief, measurement is defined as “the assignment of numerals to objects and events according to rules” (Stevens, 1946, p.340).

One way to ensure the quality test items is by using item analysis, including item difficulty and Differential Item Functioning (DIF). It is essential to note which items are most challenging for students by taking into account the students' ability and item difficulty instead of the students' raw total score. Differential Item Functioning (DIF) seeks whether an item favoring a certain group of students, such as gender, school location, socioeconomic status, etc. Such analysis is vital in constructing fair and appropriate items for assessment. The item analysis can be conducted by employing Rasch analysis that enables researchers or educators to not only construct the instruments but also get more insight to modify the items based on students' proficiency development (Boone, 2016).

There are many studies conducted concerning the instrument development, employing Rasch analysis, including test validation (such as Catley et al., 2013; Christensen et al., 2019; da Rocha et al., 2013; Franchignoni et al., 2013; Lin et al., 2018). However, most of them are in the field of psychology or medicines. Altough, some studies are in mathematics education, such as (Bansilal et al., 2019); (Ling et al., 2018); (Mirza & Hussain, 2018);(Wijsman et al., 2016), the application of Rasch analysis for Indonesian context is still limited. Thus, this study aimed to provide a preliminary insight, to the application of Rasch analysis in test item development, specifically for item analysis and item bias examination so that this technique becomes more widely used in the field of mathematics education.

Rasch Analysis

Rasch (1960), in his book entitled “Probabilistic models for some intelligence and attainment tests”, elaborated the main principle of Rasch model as “a person having greater ability than another person should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another means that for any person the probability of solving the second item is the greater one” (Rasch, 1960, p.117). The foundation of Rasch model is addressing the simple question of “when a person with this ability (number of item correct) encounters an item of this difficulty (number of persons who succeeded on the item), what is the likelihood that this person gets this item correct?” (Bond & Fox, 2016, p.11). Rasch model solves this question by concluding that the likelihood of someone's success in solving the item is relying on his/her ability and the item's difficulty. The Rasch model integrates both models of ordering the students based on their ability and the problems (items) based on their difficulty.

Originally Rasch model was created for the dichotomous item, items having only two possible responses (e.g., yes or no; correct or incorrect) (Rasch, 1960). It was later developed for polytomous items, items with more than two possible responses (e.g., Likert-scale items) (Andrich, 1978). During the time, the model is developed from the rating scale model to the partial credit model, allowing items with a partial score (e.g., mathematics problems allowing partial score for incomplete solution) to be analysed employing the Rasch analysis.

METHOD

Participants

The participants of the study were 48 final year students enrolling in one of the mathematics education department in Syiah Kuala University, Banda Aceh, Indonesia. The participants consisted of 42 females (87.5%) and 6 males (12.5%). The participants take the comprehensive mathematics test that is compulsory for all mathematics education students before proceeding to complete their mathematics education degree.

Data Analysis

The dataset used for the study were taken from one cycle of mathematics comprehensive test involving 48 students of undergraduate mathematics education students. The test, with a short answer, comprised 30 items related to algebra (item 1-8), calculus (item 9-16), geometry (item 17-23), and statistics (24-30). The test was developed based on the basic competences that should be mastered by the students. Face validity was done by some mathematics lecturers to assess the items and revision was made accordingly. Rasch analysis is used to assess the item difficulty and differential item functioning (DIF). There is a wide range of software available for analysis, such as RUMM, Winsteps, and Conquest. This study employed the ConQuest 4 software developed by the Australian Council for Educational Research (ACER). It is a powerful tool enabling researchers to investigate the properties of performance assessments, traditional assessment, rating scales, and partial credit. The data was analysed by Rasch analysis with a partial credit model as the partial score was given to incomplete or partially correct items.

RESULT AND DISCUSSION

Item Difficulty

Table 1 presents the response model parameter estimates for item difficulty of each item in the test. Rasch analysis expresses the person ability and item difficulty using a logit, natural logarithm of the likelihood for someone to be able to solve a problem or task. The magnitude of the logits indicates the item. The greater the logit means, the more difficult the item for the participants difficulty, whether it is easier or more difficult. Table 1 shows that the item difficulty for algebra ranges between -1.340 and 1.435, indicating that item 2 is the easiest and item 6 as the most difficult item. The range of item difficulty for calculus items is from 0.052 to 1.525, indicating that item 11 is the most difficult item among the calculus items. Compared to the algebra items, the item difficulty estimates, as shown by the logit, show that the participants found that the algebra items are easier than the calculus items. Furthermore, Table 1 also shows the item difficulty for geometry and statistics items. The ranges are from -1.960 to 1.230 and -1.295 to 0.857 for geometry and statistics items, respectively. The Rasch analysis also provided the separation reliability or item reliability for the test and the significant level; they are 0.783 and 0.00, respectively (df=29). The item reliability is above 0.70 and considered to be acceptable (Bond & Fox, 2016).

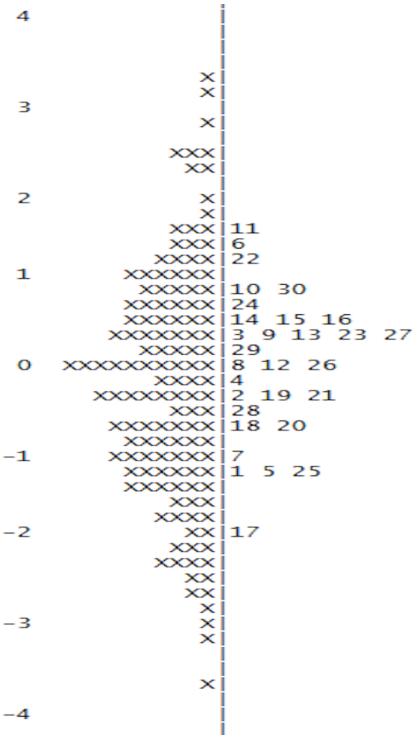
Reviewing the items in the test as a whole, it can be seen that the item difficulty ranging between -1.960 and 1.525 for all items in the test. The easiest item belongs to a geometry item, and the most difficult item is from the calculus items. Looking at the spread of item difficulty per mathematics strands, including algebra, calculus, geometry, and statistics, the estimates indicate that most easier items belong related to algebra as evidenced by negative logits. This is on the contrary to a study conducted by (Chow, 2011), revealing that students have difficulty in learning algebra. However, the contrast results may be due to the number of samples and the type of problem used in the study.

Figure 1 presents the item-person map showing the distribution of item difficulty and person ability. "Given that the mean item difficulty is arbitrarily set at 0 logits, the mean person estimate (i.e. group average) would be closer to 0 for a well-targeted test" (Bond and Fox, 2016, p.73). Thus, it can be said that Figure 1 shows that the test is relatively well mathed to the sample.

Table 1. Response Model Parameter Estimates for Item Difficulty

Item	Topics	Estimate	Std Error
1	Algebra	-1.340	0.411
2	Algebra	-0.361	0.308
3	Algebra	0.300	0.304
4	Algebra	-0.118	0.379
5	Algebra	-1.189	0.421
6	Algebra	1.435	0.634
7	Algebra	-1.002	0.405
8	Algebra	0.046	0.383
9	Calculus	0.342	0.382
10	Calculus	0.849	0.401

Item	Topics	Estimate	Std Error
11	Calculus	1.525	0.641
12	Calculus	0.052	0.373
13	Calculus	0.309	0.320
14	Calculus	0.455	0.382
15	Calculus	0.572	0.444
16	Calculus	0.544	0.382
17	Geometry	-1.960	0.466
18	Geometry	-0.746	0.402
19	Geometry	-0301	0.383
20	Geometry	0.309	0.402
21	Geometry	-0.360	0.384
22	Geometry	1.203	0.393
23	Geometry	0.322	0.384
24	Statistics	0.731	0.384
25	Statistics	-1.295	0.410
26	Statistics	-0.001	0.380
27	Statistics	0.257	0.384
28	Statistics	-0.487	0.385
29	Statistics	0.105	0.382
30	Statistics	0.857	0.390



```
  4              |
                 |
                 |
                 |
                X|
                X|
  3              |
                X|
                 |
              XXX|
               XX|
                 |
  2             X|
                X|
              XXX|11
              XXX|6
             XXXX|22
  1        XXXXXX|
            XXXXX|10  30
           XXXXXX|24
           XXXXXX|14  15  16
          XXXXXXX|3  9  13  23  27
            XXXXX|29
  0    XXXXXXXXXX|8  12  26
             XXXX|4
         XXXXXXXX|2  19  21
              XXX|28
          XXXXXXX|18  20
           XXXXXX|
 -1       XXXXXXX|7
           XXXXXX|1  5  25
           XXXXXX|
              XXX|
             XXXX|
 -2            XX|17
              XXX|
             XXXX|
               XX|
               XX|
                X|
 -3             X|
                X|
                 |
                 |
                X|
                 |
 -4              |
                 |
```

Figure 1. The Item-Person Map

Differential Item Functioning (DIF)

An item is said to exhibit Differential Item Functioning (DIF) "if the response probabilities for that item cannot be fully explained by the ability of the student and a fixed set of difficulty parameters for that item" (Adams & Wu, 2010, p.1). ConQuest enables the examination of any DIF indication through its multi-faceted modelling capabilities, allowing interaction between facets. The ACER ConQuest 4 can detect DIF statistically. Uniform DIF detected when the main effect of the person is significant. In this study, Differential Item Functioning (DIF) was measured using Rasch analysis by employing ACER ConQuest 4 software to investigate whether the items in the test are favouring the same way for both male and female students. The DIF investigation between a group of the sample, in this case, gender, is paramount in each process of an instrument or test development as the test developers should ensure that the tool used for the measurement does not favour a certain group. A test should be inclusive to everyone.

Table 2. Response Model Parameter Estimates for Item*Gender

Item	Topics	Estimate	Gender	Estimate	Gender
1	Algebra	0.313	Male	-0.313	Female
2	Algebra	0.103	Male	-0.103	Female
3	Algebra	0.704	Male	-0.704	Female
4	Algebra	0.225	Male	-0.225	Female
5	Algebra	-0.280	Male	0.280	Female
6	Algebra	0.281	Male	-0.281	Female
7	Algebra	-0.027	Male	0.027	Female
8	Algebra	0.449	Male	-0.449	Female
9	Calculus	0.155	Male	-0.155	Female
10	Calculus	0.072	Male	-0.072	Female
11	Calculus	0.235	Male	-0.235	Female
12	Calculus	-1.305	Male	1.305	Female
13	Calculus	-0.780	Male	0.780	Female
14	Calculus	0.035	Male	-0.035	Female
15	Calculus	0.854	Male	-0.854	Female
16	Calculus	-0.052	Male	0.052	Female
17	Geometry	-0.030	Male	-0.030	Female
18	Geometry	-0.283	Male	-0.283	Female
19	Geometry	0.042	Male	0.030	Female
20	Geometry	-0.282	Male	0.282	Female
21	Geometry	0.104	Male	-0.104	Female
22	Geometry	-0.705	Male	0.705	Female
23	Geometry	-0.577	Male	0.577	Female
24	Statistics	-0.231	Male	0.231	Female
25	Statistics	0.273	Male	- 0.273	Female
26	Statistics	0.976	Male	-0.976	Female
27	Statistics	-0.519	Male	0.519	Female
28	Statistics	0.226	Male	-0.226	Female
29	Statistics	-0.388	Male	-0.388	Female
30	Statistics	-0.364	Male	0.364	Female

The Rasch analysis run for DIF also provides information related to the difference in performance between male and female students. The estimates for male is -0.490 and estimates for female is 0.490. These *estimates* indicates that the *estimates* of male students' score is 0.98 lower than females, and the difference is significant (p=0.000). Table 2 depicts the response model parameter estimates for item*gender. It shows the *estimates* of the difference in item difficulty between males and females. The estimates show that items behave differently between male and female students. However, the results show that it is not significant (p=0.951, df=29, $\chi 2$=17.64). Since it is not significant, it can be concluded that no DIF is detected for the test items, indicating that items do not favour either male or female.

The separation reliability of 0.40 indicates weak reliability (Bond & Fox, 2016).

CONCLUSION

The Rasch analysis, employing ACER ConQuest 4 software, provides some insight related to the item difficulty and DIF of the test. The item difficulty estimates and the item-person map depict clear information related to the item difficulty, with item 17 related to the geometry being the easiest (-1.960) and item 11 related to calculus (1.525) being the most difficult for students. However, there is no indication of DIF as the parameter estimates for item*gender are not significant, indicating that the items are not biased toward male or female students. The results of the Rasch analysis and its elaboration in the discussion section conclude that Rasch analysis can be applied for item analysis (such as: the easiest and hardest items) and item bias analysis (such as whether items performing differently between gender). It is expected that this analysis can be widely used in other instrument development in the field of mathematics education.

However, there is a limitation within this preliminary study; the sample was only 48 students, fairly small to produce a rigorous and reliable result. Nevertheless, this finding provides a useful insight into the bigger study conducted in the future. Further study should be undertaken for more reliable and rigorous findings to overcome the limitation of this study related to sample size and the unbalance number between male and female students. Despite its limitation, this study has been successful in presenting an introduction of Rasch analysis for wider mathematics education research, for its wide range of applications including items/instrument development.

REFERENCES

Adams, R., & Wu, M. (2010). *Differential Item Functioning*. ACER. www.acer.org

Andrich, D. (1978). A Rating Formulation for Ordered Response Categories. *Psychometrika*, *43*(4), 561–573.

Bansilal, S., Long, C., & Juan, A. (2019). Lucky Guess? Applying Rasch Measurement Theory to Grade 5 South African Mathematics Achievement Data. *Journal of Applied Measurement*, *20*(2), 206–220.

Black, P., Harrison, C., & Lee, C. (2003). *Assessment for Learning: Putting It into Practice*.

McGraw-Hill Education.

Bond, T. G., & Fox, C. M. (2016). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE—Life Sciences Education*, *15*(4), 1–7.

Boud, D., & Falchikov, N. (2007). *Rethinking Assessment in Higher Education: Learning for the Longer Term*. Routledge.

Catley, M. J., O'Connell, N. E., & Moseley, G. L. (2013). How Good is the Neurophysiology of Pain Questionnaire? A Rasch Analysis of Psychometric Properties. *The Journal of Pain*, *14*(8), 818–827.

Chow, T.-C. F. (2011). *Students' Difficulties, Conceptions and Attitudes Towards Learning Algebra: An Intervention Study to Improve Teaching and Learning*. Curtin University.

Christensen, K. S., Oernboel, E., Nielsen, M. G., & Bech, P. (2019). Diagnosing Depression in Primary Care: A Rasch Analysis of the Major Depression Inventory. *Scandinavian Journal of Primary Health Care*, *37*(1), 105–112.

da Rocha, N. S., Chachamovich, E., de Almeida Fleck, M. P., & Tennant, A. (2013). An Introduction to Rasch Analysis for Psychiatric Practice and Research. *Journal of Psychiatric Research*, *47*(2), 141–148.

Franchignoni, F., Mora, G., Giordano, A., Volanti, P., & Chiò, A. (2013). Evidence of Multidimensionality in the ALSFRS-R Scale: A Critical Appraisal on Its Measurement Properties Using Rasch Analysis. *Journal of Neurology, Neurosurgery & Psychiatry*, *84*(12), 1340–1345.

Gardner, J. (2012). *Assessment and Learning*. Sage.

Guler, N., Uyanik, G. K., & Teker, G. T. (2014). Comparison of Classical Test Theory and Item Response Theory in Terms of Item Parameters. *European Journal of Research on*

Education, *2*(1), 1–6.

Lin, C.-Y., Pakpour, A. H., Broström, A., Fridlund, B., Årestedt, K., Strömberg, A., Jaarsma, T., & Mårtensson, J. (2018). Psychometric Properties of the 9-Item European Heart Failure Self-Care Behavior Scale Using Confirmatory Factor Analysis and Rasch Analysis Among Iranian Patients. *Journal of Cardiovascular Nursing*, *33*(3), 281–288.

Ling, M.-T., Pang, V., & Ompok, C. C. (2018). Measuring Change in Early Mathematics Ability of Children Who Learn Using Games: Stacked Analysis in Rasch Measurement. *Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings: Rasch and the Future*, 215–226.

Mirza, A., & Hussain, N. (2018). Performing Below the Targeted Level: An Investigation into KS3 Pupils’ Attitudes Towards Mathematics. *Journal of Education and Educational Development*, *5*(1), 1–17.

Rasch, G. (1960). *Probabilistic models for some intelligence tests and attainment tests*. Danmarks Paedagogiske Institut.

Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, *103*, 677–680.

Wijsman, L. A., Warrens, M. J., Saab, N., van Driel, J. H., & Westenberg, P. M. (2016). Declining Trends in Student Performance in Lower Secondary Education. *European Journal of Psychology of Education*, *31*(4), 595–612.