Vol. 4, No. 2 (2020) 197-203 ISSN: 2597-4866

Indonesian Journal of Primary Education

Disclosure of Difficulty Distribution of HOTS-Based Test Questions through Rasch Modeling

Ade Yulianto[1], Ari Widodo[2]

[1.2]Master of Basic Education Study Program, Graduate School of the Indonesian University of Education, Bandung.
*Corresponding author: adeyulianto@upi.edu[1] , widodo@upi.ed[2]

Received November 8, 2020; Revised November 16, 2020; Accepted 10 December 2020
Published 23 December 2020

Abstract

Many researchers or teachers need more and more complete information about the evaluation of the implementation of learning or the ability measurement tools used to find out how much learning outcomes are shown by their students. In this article, it is shown how to analyze questions based on the level of difficulty and suitability of test item items; especially on HOTS (Higher-Order Thinking Skill) based test questions which were developed based on a cognitive hierarchy adopted from Bloom's taxonomy (C4, C5, & C6). The analysis process was carried out based on Sumintono's (2015) explanation of the combination of standard deviation (SD) values and logit average values (Mean). Then perform the criteria for the outfit mean square (MNSQ) value, the Z-standard outfit value (ZSTD) and the point measure correlation value (Pt. Measure Corr) (Boone et al., 2014). The analysis technique was carried out through Rasch modeling assisted by the Winsteps 3.75 application. As for the grouping of difficulty levels according to Sumintono (2015), namely 1) difficult question category (logit value is greater + 1SD); 2) difficult question category (value 0.0 logit +1 SD); 3) easy question category (value 0.0 logit -1 SD); and 4) the category of questions is very easy (value less than -SD), as well as for the criteria used to measure the suitability of item items using, 1) the value of 0.5 <MNSQ <1.5; 2) value -2.0 <ZSTD <+2.0; and 3) the value of 0.4 <PT-Measure Corr <0.85, (Boone et al., 2014). The results of the analysis show that there are variations in the level of difficulty and suitability of HOTS-based test items. The item analyzed had an acceptable level of suitability and was feasible to maintain because all items met these three criteria. Thus, the collection of HOTS-based test questions is in a good category because it can identify students' various abilities in higher-order thinking with varying levels of difficulty.

Keywords:*Analysis of Item Difficulty and Suitability; HOTS-Based Test Questions; Rasch Modeling.*

Abstract

Many researchers or teachers need more and more complete information about evaluating the implementation of learning or measuring ability tools used to find out how much learning outcomes are shown by their students. In this article, it is shown how to analyze questions based on grouping the level of difficulty and suitability of test item items; especially on the HOTS (Higher-Order Thinking Skill) based test questions which were developed based on the cognitive hierarchy adopted from Bloom's taxonomy (C4, C5, & C6). The analysis process was carried out based on Sumintono's (2015) explanation about the combination of the standard deviation (SD) value and the logit mean (Mean). Then perform the criteria for outfit mean square (MNSQ), outfit Z-standard (ZSTD) and point measure correlation (Pt. Measure Corr) values (Boone et al., 2014). The analysis technique was carried out through Rasch modeling assisted by the Winsteps 3.75 application. The difficulty level grouping according to Sumintono (2015) is 1) difficult question category (logit value is greater than +1 SD); 2) the category of difficult questions (value 0.0 logit +1 SD); 3) easy question category (value 0.0 logit -1 SD); and 4) the category of questions is very easy (value is smaller than –SD), as well as for the criteria used to measure the suitability of the item items using, 1) the value of 0.5 < MNSQ < 1.5; 2) value -2.0 < ZSTD < +2.0; and 3) the value of 0.4 < PT-Measure Corr < 0.85, (Boone et al., 2014). The results of the analysis show that there is a diversity of levels of difficulty and suitability of HOTS-based test items. The item items analyzed have an acceptable level of conformity and deserves to be maintained because all item items meet the three criteria. Thus, the collection of HOTS-based test questions is categorized as good because it can identify various students' abilities in higher-order thinking with varying levels of difficulty and this has implications for the teacher's technique in presenting quality written test questions.

Keywords:*Difficulty and Item Suitability Analysis; HOTS-Based Test Questions; Rasch modeling.*

PRELIMINARY

In the learning process, of course, has a learning goal regarding a number of knowledge or skills that must be achieved by students. In an effort to find out the process of achieving this, a measuring instrument is needed to determine the value, score or percentage achieved by students related to the learning objectives. In the process or end of learning, it is necessary to measure the process and learning outcomes in the form of numbers that reflect the achievements of the learning process and results. According to Mardapi (2017), measurement is basically an activity of determining numbers on an object systematically. Measurement is a process that describes student performance using a quantitative scale (number system) in such a way that the qualitative nature of the student's performance is expressed in numbers (Alwasillah et al, 1996). Thus, measurement in education means measuring student attributes or characteristics (Safithry, 2019).

Measurement in education is closely related to tests. This is because one way that is often used to measure the results that have been achieved by students is by testing (Tillar, 2018). In the measurement process, of course, the teacher must use measuring tools, either test or non-test (Sadikin, 2018). According to Zainul (2001), the test is defined as a question or task or a set of tasks that are planned to obtain information about an educational attribute. The test used is adjusted to the subject or field of science that will be used as a test source. The test questions were developed to explore students' higher-order thinking abilities. Higher order thinking skills need to be developed in education in elementary schools as an effort to improve students' ability to think logically in answering more complicated questions and or solving more complicated problem cases. Heong (2011) explains that higher order thinking skills require a person to apply new information or prior knowledge and manipulate information to reach possible answers in new situations. The term higher-order thinking skills (HOTS) is part of Bloom's classification of thinking skills. Bloom (in Anderson & Karthwohl, 2001) classifies thinking skills into six levels, namely remembering/C-1, understanding/C-2, applying/C-3, analyzing/C-4, evaluate/ C-5 and create/C-6. Brookhart (2010) states that the levels of thinking included in the HOTS include the three highest thinking skills in Bloom's taxonomy, namely analyzing, evaluating and creating.

The results obtained from the test are the data to be analyzed with the help of the Winsteps 3.75 application through the Rasch rating scale model. The Rasch model is a well-studied measurement approach that models the relationship between item difficulty, person ability, and the probability of a given response (Andrich, 1981). The Rasch model which provides psychometric analysis techniques can be used by teachers to develop test items as well as important tools that can provide relevant information related to student assessment for learning (Suminton, 2018). The analysis of this test instrument using the Rasch model is included in the response item measurement theory. This measurement describes the interaction between the subject and the test item. This will make measurements have more precise and objective results (Sumintono & Widhiarso, 2014). Meanwhile, according to Brogde (1977), The Rasch model is usually applied to the measurement of items and subjects of persons in this context that is discussed in relation to the law of comparative judgment and other additional combined measures. According to Masters (1982), Rasch modeling can be used for various observational formats, including models for calculation analysis, repeated experiments and rating scales. In addition, the statistical picture of Rasch conformity can provide a useful framework for testing the correctness of a person's response, measuring the estimate of a person's response ability and being able to detect various disturbances to a person's response (Smith, 1986). It should be noted that the Rasch model is written as a model of the probability of an individual's response to an item and is therefore not explicitly a model of the response itself (Brogden, 1977). Georg Rasch developed an analytical model of item response theory (or Item Response Theory, IRT) in the 1960s, which was later popularized by Ben Wright (Misbach & Sumintono, 2014). With raw data in the form of dichotomous data (in the form of true and false) indicating students' abilities, Rasch formulated this into a model that relates students and items (Sumintono & Widhiarso, 2015). In addition to dichotomous data, the Rasch model can also perform analysis for polytomy data such as those developed by Andrich, which are still based on two basic theorems, the level of a person's ability and the level of item difficulty. With raw data in the form of dichotomous data (in the form of true and false) indicating students' abilities, Rasch formulated this into a model that relates students and items (Sumintono & Widhiarso, 2015). In addition to dichotomous data, the Rasch model can also perform analysis for polytomy data such as those developed by Andrich, which are still based on two basic theorems, the level of a person's ability and the level of item difficulty. With raw data in the form of dichotomous data (in the form of true and false) indicating students' abilities, Rasch formulated this into a model that relates students and items (Sumintono & Widhiarso, 2015). In addition to dichotomous data, the Rasch model can also perform analysis for polytomy data such as those developed by Andrich, which are still based on two basic theorems, the level of a person's ability and the level of item difficulty.

The analysis process through Rasch modeling in previous research was carried out in the development of character and anxiety instruments in elementary schools, so in this study it becomes a reference in the technical process of data processing (Nur et al., 2020; Karlimah et al., 2020) as well as in measuring thinking skills. critical students in elementary schools (Hamdu, et al., 2020). Rasch's model assumes that item difficulty is a trait that is influenced by the respondent's answer, and a person's ability is a trait that is influenced by an estimate of item difficulty (Linacre, 1999). The advantage of the Rasch model when compared to classical theory is that it can identify incorrect answers from experts, identify incorrect judgments, and predict missing data based on systematic response patterns (Pratama, 2020; Novinda et al, 2019).

RESEARCH METHODS

In the data collection process, 11 HOTS-based test questions were tested to be done on 20 elementary school students who had previously carried out learning on theme 2 "Always Save Energy", Sub-theme "Benefits of Alternative Energy" which was in accordance with the material in the questions developed. After that, an analysis related to the difficulty and suitability was carried out through Rasch modeling assisted by the Winsteps 3.75 application. The stages of the analysis process are through Rasch modeling with the adoption of stages carried out by Hamdu, et al. (2020), namely:

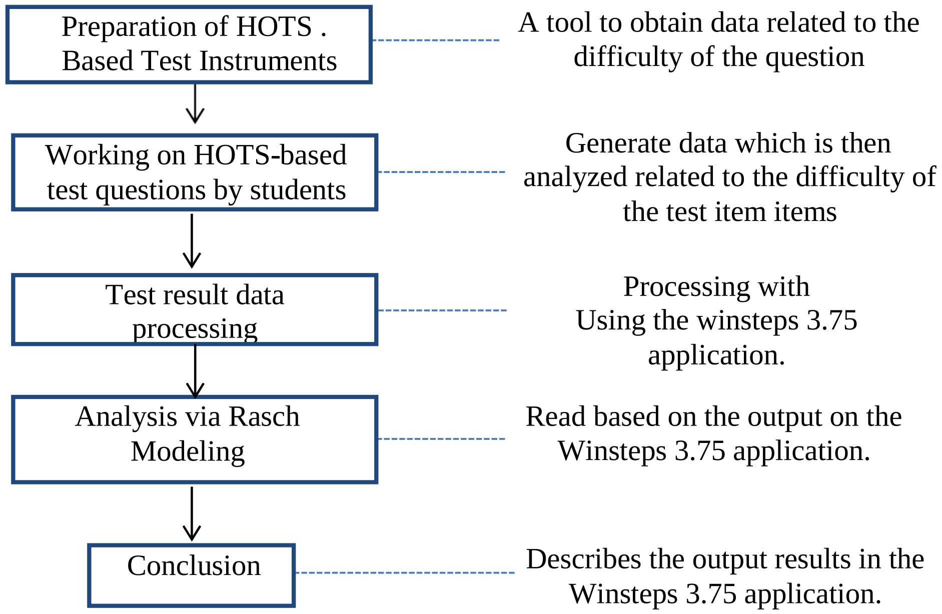

Preparation of HOTS . Based Test Instruments
A tool to obtain data related to the difficulty of the question
Working on HOTS-based test questions by students
Generate data which is then analyzed related to the difficulty of the test item items
Test result data processing
Processing with Using the winsteps 3.75 application.
Analysis via Rasch Modeling
Read based on the output on the Winsteps 3.75 application.
Conclusion
Describes the output results in the Winsteps 3.75 application.

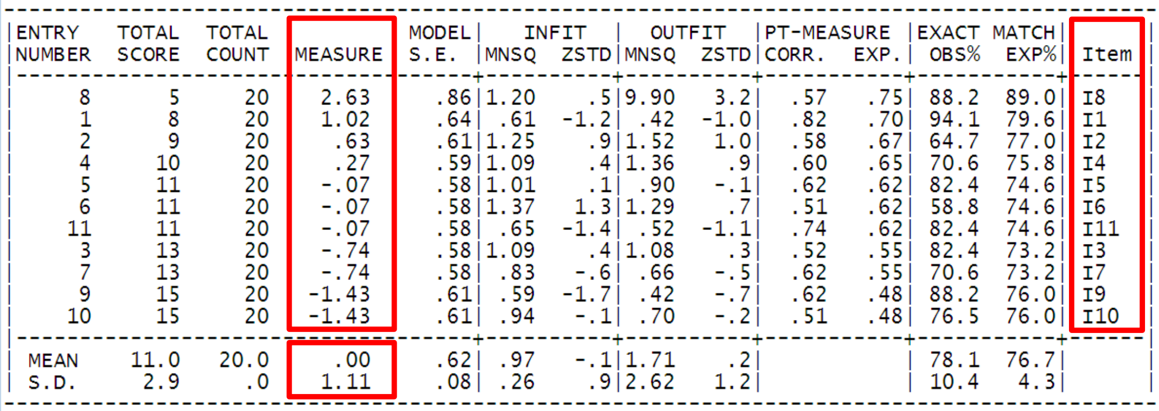
Image 1
Research Stages

RESULTS AND DISCUSSION

Analysis of the Difficulty Level of HOTS Based Test Questions

Table 1 describes several columns that can provide information about the level of difficulty of each item on the HOTS-based test. Grouping the level of difficulty of the items through a combination of the standard deviation (SD) value and the logit mean value (Sumintono, 2015). The grouping is the category of difficult questions with a logit value greater than +1 SD; the category of difficult questions with a value of 0.0 logit +1 SD; easy question category with a value of 0.0 logit -1 SD; very easy question category with a value less than –SD.

Table 1
Item Measure Output Results



```
-------------------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL           MODEL|   INFIT  |  OUTFIT  |PT-MEASURE |EXACT  MATCH|       |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ  ZSTD|MNSQ  ZSTD|CORR.  EXP.| OBS%  EXP%|  Item |
|------------------------------------+----------+----------+-----------+-----------+-------|
|     8      5     20     2.63   .86|1.20    .5|9.90   3.2|  .57   .75| 88.2  89.0| I8    |
|     1      8     20     1.02   .64| .61  -1.2| .42  -1.0|  .82   .70| 94.1  79.6| I1    |
|     2      9     20      .63   .61|1.25    .9|1.52   1.0|  .58   .67| 64.7  77.0| I2    |
|     4     10     20      .27   .59|1.09    .4|1.36    .9|  .60   .65| 70.6  75.8| I4    |
|     5     11     20     -.07   .58|1.01    .1| .90   -.1|  .62   .62| 82.4  74.6| I5    |
|     6     11     20     -.07   .58|1.37   1.3|1.29    .7|  .51   .62| 58.8  74.6| I6    |
|    11     11     20     -.07   .58| .65  -1.4| .52  -1.1|  .74   .62| 82.4  74.6| I11   |
|     3     13     20     -.74   .58|1.09    .4|1.08    .3|  .52   .55| 82.4  73.2| I3    |
|     7     13     20     -.74   .58| .83   -.6| .66   -.5|  .62   .55| 70.6  73.2| I7    |
|     9     15     20    -1.43   .61| .59  -1.7| .42   -.7|  .62   .48| 88.2  76.0| I9    |
|    10     15     20    -1.43   .61| .94   -.1| .70   -.2|  .51   .48| 76.5  76.0| I10   |
|------------------------------------+----------+----------+-----------+-----------+-------|
| MEAN    11.0   20.0      .00   .62| .97   -.1|1.71    .2|           | 78.1  76.7|       |
| S.D.     2.9     .0     1.11   .08| .26    .9|2.62   1.2|           | 10.4   4.3|       |
-------------------------------------------------------------------------------------------------------
```

Based on Table 1 above, the HOTS results can be grouped as follows:

1. The group of difficult items, namely question no. I8.
2. The group of difficult items, namely question no. I1, I2, and I4.
3. The group of easy items, namely question no. I5, I6, I11, I3, and I7.
4. The group of items is very easy, namely question no. I9 and I10.

Table 2
Difficulty Levels Based on Cognitive Level

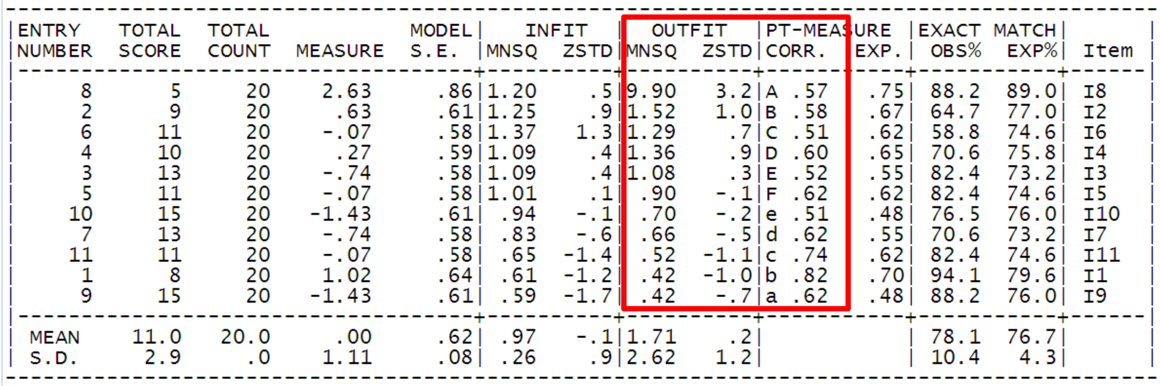
No. Question	Difficulty Level	Cognitive Level
1	Hard	C4
2	Hard	C4
3	Easy	C4
4	Hard	C4
5	Easy	C6
6	Easy	C5
7	Easy	C4
8	Difficult	C4
9	Very easy	C5
10	Very easy	C5
11	Easy	C4

Analysis of the Level of Conformity for HOTS-Based Test Questions

Furthermore, after analyzing the level of difficulty of the HOTS-based test questions, further analysis of the level of suitability of the difficulty of the items (Item Fit) can be seen using three criteria, namely the means-square outfit value (Outfit MNSQ), Outfit Z-Standard (Outfit ZSTD), and Point Measure Correlation (PT-Measure Corr) (Bonee et al., 2014; Bond & Fox, 2015; Sumintono & Widhiarso, 2015). The criteria used to check the suitability of items that are not appropriate (outliers or misfits), namely:

1. Outfit means-square value (Outfit MNSQ) received: 0.5 < MNSQ < 1.5.
2. Value of Outfit Z-Standard (Outfit ZSTD) accepted: -2.0 < ZSTD < +2.0.
3. Point Measure Correlation (PT-Measure Corr) value: 0.4 < PT-Measure Corr < 0.85.

Table 3
Item Fit Order Output Results



```
---------------------------------------------------------------------------------------------
|ENTRY    TOTAL   TOTAL               MODEL|   INFIT    |  OUTFIT   |PT-MEASURE |EXACT  MATCH|      |
|NUMBER   SCORE   COUNT    MEASURE   S.E. |MNSQ   ZSTD |MNSQ   ZSTD|CORR.  EXP.| OBS%   EXP%| Item |
|------------------------------------+------------+-----------+-----------+-----------+------|
|     8      5      20      2.63      .86|1.20     .5|9.90    3.2|A .57   .75| 88.2   89.0| I8   |
|     2      9      20       .63      .61|1.25     .9|1.52    1.0|B .58   .67| 64.7   77.0| I2   |
|     6     11      20      -.07      .58|1.37    1.3|1.29     .7|C .51   .62| 58.8   74.6| I6   |
|     4     10      20       .27      .59|1.09     .4|1.36     .9|D .60   .65| 70.6   75.8| I4   |
|     3     13      20      -.74      .58|1.09     .4|1.08     .3|E .52   .55| 82.4   73.2| I3   |
|     5     11      20      -.07      .58|1.01     .1| .90    -.1|F .62   .62| 82.4   74.6| I5   |
|    10     15      20     -1.43      .61| .94    -.1| .70    -.2|e .51   .48| 76.5   76.0| I10  |
|     7     13      20      -.74      .58| .83    -.6| .66    -.5|d .62   .55| 70.6   73.2| I7   |
|    11     11      20      -.07      .58| .65   -1.4| .52   -1.1|c .74   .62| 82.4   74.6| I11  |
|     1      8      20      1.02      .64| .61   -1.2| .42   -1.0|b .82   .70| 94.1   79.6| I1   |
|     9     15      20     -1.43      .61| .59   -1.7| .42    -.7|a .62   .48| 88.2   76.0| I9   |
|------------------------------------+------------+-----------+-----------+-----------+------|
| MEAN     11.0    20.0       .00      .62| .97    -.1|1.71     .2|           | 78.1   76.7|      |
| S.D.      2.9      .0      1.11      .08| .26     .9|2.62    1.2|           | 10.4    4.3|      |
---------------------------------------------------------------------------------------------
```

From Table 3 above, it is found that item I8 does not meet the MNSQ value; item I8 does not meet the ZSTD value; and all items meet the PT-Measure Corr score. If the items in the three criteria (MNSQ, ZSTD, and Pt. Measure Corr) are not met, it can be ascertained that the questions are not good enough so they need to be repaired or replaced (Bonee et al., 2014; Bond & Fox, 2015). From this quote, all HOTS-based test items analyzed have an acceptable level of conformity and deserves to be maintained because there are no items that do not meet the three criteria. This reasoning can indicate that there is a diversity of levels of difficulty of the appropriate questions in table 1. A good collection of test questions can identify various abilities of students with varying levels of difficulty. If the level of difficulty in the collection of test questions is high, it can be ascertained that students cannot answer correctly or do not understand the questions given. However, if the level of difficulty in the set of questions is low, it can be ascertained that many students can answer correctly or easily.

CONCLUSION

Based on the results of the analysis and discussion, it can be concluded that a number of test questions show a diversity of levels of difficulty with an acceptable level of conformity. The difficulty level is related to the fractional number material, namely, 1) the majority of difficult item groups are found in HOTS-based test questions at the C4 cognitive level; 2) the majority of difficult item groups are found in HOTS-based test questions at the C4 cognitive level; 3) the majority of easy item groups are found in HOTS-based test questions at various cognitive levels, namely C4, C5 and C6; 4) the group of items is very easy, the majority are in the HOTS-based test questions at the C5 cognitive level.

The results of the analysis through Rasch modeling are more specifically to provide a comprehensive picture of learning on theme 2 "Always Save Energy", Sub-theme "Benefits of Alternative Energy" which is carried out on a certain subject and time. The results of the analysis of this Rasch modeling may be different or the same by taking into account the conditions and learning situations, such as relating to the characteristics of students and implementers of learning in certain classes or schools. However, the analysis process of Rasch modeling can be used by teachers in schools to be able to comprehensively identify the learning process and make test questions.

BIBLIOGRAPHY

Alhamdulillah, et al. (1996). Glossary of educational Assessment Terms. Jakarta: Ministry of Education and Culture.

Anderson & Krathwohl. (2001). A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives. New York: Addison Wesley Longman, Inc

Andrich, D. (1981). Book Review : Probabilistic Models for Some Intelligence and Attainment Tests (expanded edition. Applied Psychological Measurement, 5(4), 545– 550. doi:10.1177/014662168100500413.

Bond, TG & Fox, CM (2015). Applying the Rasch Model, Fundamental Measurement in the Human Science (3rd edition). New York: Routledge.

Bonee, WJ, Staver, JR, & Yale, MS (2014). Rasch Analysis in the Human Science. Dordrecht: Springer

Brogden, HE (1977). The rasch model, the law of comparative judgment and additive conjoint measurement. Psychometrics, 42(4), 631–634. doi:10.1007/bf02295985.

Brookhart, S. M. (2010). Higher-Order Thinking In Your Classroom. ACDC, Alexandria, Virginia USA.

Heong, Y. et.al. (2011). The level of marzano higher order thinking skills among technical education students. International Journal of Social Science and Humanity, 1(2), 121.

Karlimah, K., Andriani, D., & Suryana, D. (2020). Development of Mathematical Anxiety Instruments with a Rasch Model Analysis. The Open Psychology Journal, 13(1).

Linacre, JM (1999). Investigating Rating Scale Category Utility. Journal of Outcome Measurement. 3(2), 103-122.

Mardapi, D. (2017). Educational Measurement, Assessment, and Evaluation. Yogyakarta: Parama Publishing

Masters, GN (1982). A Rasch Model For Partial Credit Scoring. Psychometrics, 47(2), 149–174. doi:10.1007/bf02296272.

Misbach, IH and Sumintono, B. (2014) Development and validation of the instrument "student perceptions of the teacher's moral character" in Indonesia using the Rasch Model. In: National Psychometric Seminar: "Development of Valid Character Assessment Instruments", 24 May 2014, Solo, Indonesia.

Novinda, MRR, Silitonga, HTM, & Hamdani, H. (2019). Development of Multiple Choice Test Using Rasch Model for Class X Pontianak. Journal of Equatorial Education and Learning, 8(6).

Nur, L., Conscience, LA, Suryana, D., & Ahmad, A. (2020). Rasch model application on character development instrument for elementary school students. International Journal of Learning, Teaching and Educational Research, 19(3), 437-459.

Pratama, D. (2020). Quality Analysis of Teacher-Made Tests Through Rasch's Item Response Theory (IRT) Approach. Tarbawy: Journal of Islamic Education, 7(1), 61-70.

Sadikin, A. (2018). Evaluation and Process Biology learning outcomes. Jambi: Salim Media Indonesia.

Saftihry, EA (2019). Technical Assessment of Tests and Nontests. Malang: CV. IRDH.

Smith, RM (1986). Person Fit in the Rasch Model. Educational and Psychological Measurement, 46(2), 359–372. doi:10.1177/001316448604600210.

Sumintono, B. (2018). Rasch Model Measurements as Tools in Assessment for Learning. In 1st International Conference on Education Innovation (ICEI 2017). Atlantis Press.

Sumintono, B., & Widhiarso, W. (2015). RASCH Modeling Application in Educational Assessment. In TrimKom Publishing Home.

Tilaar, ALF (2018). Mathematics Learning Assessment. Tondano: Unima Press.

Zainul, A. (2001). Alternative assessment. Jakarta: Director General of Higher Education.