**Integrated Science Education Journal**
Vol. 4, No. 2, May 2023, pp. 62~68
ISSN: 2716-3725, DOI: 10.37251/isej.v4i2.325


# The Rasch Model: Implementation of Physics Learning Evaluation Instrument Based on Higher Order Thinking Skills

**Rika Irmayanti[1], Muhammad Rusdi[2], Yusnaidar[3]**
[1,2,3] Science Education, Pascasarjana, Universitas Jambi, Jambi, Indonesia

**Article Info**

***Article history:***

Received May 1, 2023
Revised May 15, 2023
Accepted May 26, 2023
OnlineFirst May 30, 2023

***Keywords:***

HOTS
Physics Learning
Rasch Model

**ABSTRACT**

**Purpose of the study:** This research was conducted to analyze the test instruments used to measure students' HOTS abilities on static fluid material at Senior high school 2 Jambi City.

**Methodology:** The evaluation instrument provided was in the form of 20 items of two-tier multiple choice questions related to static fluid material. The Rasch model is used to get fit items. This analysis was carried out with the help of Ministeps software. Respondents in this study were 36 students of class XI MIPA 1 Senior high school 2 Jambi City.

**Main Findings:** From the output of the Ministeps program, the results obtained were 20 questions according to the Rasch model with an average Outfit MNSQ score for person and item respectively 0.94 and 0.94. While the Outfit ZSTD values for persons and items are 0 and -0.1 respectively, while the reliability of the instrument expressed in Cronbach's alpha is 0.79.

**Novelty/Originality of this study:** The evaluation instrument used for physics subjects on static fluid material is fit with the Rasch model. so that the instrument meets the criteria for use.



***Corresponding Author:***
Rika Irmayanti,
Science Education, Universitas Jambi, Jambi, Indonesia
Jl. Raden Mattaher No.21, Ps. Jambi, Jambi City, Jambi 36133, Indonesia
Email: rikairmayanti4@gmail.com

## 1. INTRODUCTION

Evaluation of learning is an activity that educators must implement in every learning implementation [1]. In every series of good learning stages, the evaluation must be carried out continuously [2]. Learning evaluation activities must be carried out in a planned and interconnected manner between educational supporting factors [3]. This is so that evaluation activities can function as a tool to find out how far the effectiveness of the implementation of teaching and learning has been carried out to achieve the goal. Teachers need evaluation instruments to conduct an assessment, test students' understanding, and determine how far the learning process has achieved results according to the goals set [4]. One of the evaluation instruments needed to determine student understanding is a cognitive assessment instrument[5]. Implementation of the available curriculum is the teacher's job to be able to guide students to be able to think critically and systematically and be able to conclude problem-solving, as well as have higher-order thinking skills (HOTS) [6]. Indicators to measure higher-order thinking skills include analyzing, evaluating, and creating [7].

The output of students with higher-order thinking skills is not only developed in the learning process but must also be supported by evaluations that reflect higher-order thinking skills because evaluation is an integral part of classroom learning [8]. Evaluation can be used to measure the success of achieving the learning indicators carried out [9]. Three aspects of learning outcomes indicators are cognitive, psychomotor, and affective. The evaluation used to measure higher-order thinking skills is based on cognitive aspects. Evaluation instruments that measure higher-order thinking skills can use various assessments, such as modified multiple choice. The two-tier multiple-choice question form is one of the Modified multiple-choice alternatives that can be used to measure higher-order thinking skills. The two-tier multiple-choice question form was developed by Treagust [10]. Treagust uses multilevel multiple-choice questions to diagnose students' ability to understand science concepts. The form question consists of two levels of questions; the first level is the content of the question, which has two alternative answers, and the second level is the reason for the answer chosen based on the first choice.

Haladyna & Downing, and Treagust stated the advantages of the two-tier multiple-choice question form, one of which is used for test purposes that measure students' cognitive abilities at a higher level (Higher Order Thinking Skills) [11]. The two-tier multiple-choice question form can be used to help test students' understanding and help identify misconceptions that students may have. The inclusion of reasons at the second level of the two-tier multiple-choice question form can be used to improve higher-order thinking skills and see students' ability to give reasons [12]. The inclusion of reasons at the second level of this question can be used to reduce the occurrence of answers based on guesswork and obtaining results based on luck which is often the weakness of the usual multiple-choice questions. Objective, easy and fast assessment of questions is the advantage of the two-tier multiple choice question compared to other questions of higher order thinking skills, for example, essay questions.

This study focuses on analyzing physics evaluation instruments on static fluid material. In this study, the evaluation was carried out using an instrument of two-tier multiple-choice questions following the HOTS questions. This instrument measures students' abilities in a static fluid material. This instrument was designed as a two-tier multiple-choice, namely multilevel multiple-choice questions consisting of the main question and the reason for the answer to the main question [13]. In this instrument, the aspect considered is students' abilities in higher-order thinking. The evaluation instrument was tested on class XI MIPA 1 students at senior high school 2 Jambi City and then analyzed using the Rasch model.

Georg Rasch developed an analytical model from grain response theory in the 1960s that is similar to what is commonly called the IRT 1PL (one parameter) [14]. This mathematical model was later popularized by Benjamin Wright Linacre [15]. With raw data in dichotomous data (true and false), which indicates student abilities, Rasch formulates this into a model that connects students and items [16]. A student who can do 80% of the questions correctly certainly has better abilities than other students who can only do 65% of the questions. The data (percentage) shows that the raw data obtained is nothing but an ordinal data type that shows rank and is not linear [17]. Because ordinal data do not have the same interval, it needs to be converted into ratio data for statistical analysis. So if someone gets a score of 80%, the odds ratio is 80:20 (meaning: 80 correct scores compared to 20 wrong scores), which is nothing but frequency/ratio comparison data that is more appropriate for measurement purposes. Through this ratio comparison data, Rasch developed a measurement model that determines the relationship between the level of student ability (person ability) and item difficulty (item difficulty) by using a logarithmic function to produce measurements at the same intervals. The result is a new logit (log odds unit) which shows the student's ability and item difficulty later from the logit value obtained [18].

The data is in the form of numbers which are raw data scores from the exams carried out by students from the exam questions/instruments given. The instrument is designed from variables that have been satisfactorily defined (e.g., quantitative ability), then the relevant constructs are identified (i.e., those that can be measured through arithmetic tests, number series, and quantitative comparisons), items are created and developed to be able to measure the variables identified meant. At the same time, the answer choices provided follow the scoring pattern adhered to by classical test theory (CTT). In the context of Rasch modeling, this 'fixed' scoring pattern is a measurement whose results depend on who is being measured (test-dependent scoring) [19]. The concept of objective measurement in the social sciences and educational assessment, according to Mok and Wright (2004), must have five criteria [20], that is, provides a linear measure with equal intervals, carry out the proper estimation process, find inappropriate items (misfits) or unusual (outliers), overcoming missing data and produces replicable measurements (independent of the parameters studied). These five conditions can only be met by Rasch modeling. This suggests that the quality of measurements in educational assessments made using Rasch modeling will have the same quality as measurements made in the physical dimension in physics (e.g., measuring length with a centimeter ruler, measuring weight with a kilogram balance, etc.).

When viewed further, the logit scale (log odds unit) produced in Rasch modeling is a scale with equal intervals and is linear in nature, derived from the ratio data (odds ratio) and not the raw score data obtained (1). Therefore, estimating a person's ability or level of difficulty with the problem will have a more precise estimation value and can be compared with each other because they have the same units (logit) (2). Since the algorithm used will sort the respondents from high to low ability, which simultaneously sorts the questions from easy to difficult, there will be inaccuracies/consistency of answers from respondents (misfit) or patterns that are out of the ordinary (outliers) easily detected, as well as for the pattern of responses received by a particular question (3). The sorting of the respondents' abilities and the difficulty of the questions in a structured way also makes the Rasch model able to make predictions when there is missing data (4). The resulting logit scale will bring up a value that depends on the response pattern, not on the initial score determined, so Rasch modeling will always produce independent measurements (5).

Analysis with Rasch modeling produces fit statistics analysis, which informs researchers on whether the data obtained ideally depicts that people with high abilities provide patterns of answers to items according to their difficulty level. The parameters used are the infit and outfit of the mean square and standardized values. Infit (inlier sensitive or information weighted fit) is the sensitivity of the pattern of response to the target item on the respondent (person) or vice versa, while outfit (outlier sensitive fit) measures the sensitivity of the pattern of response to items with a certain level of difficulty on the respondent or vice versa [21]. Quantitative research in the social sciences and educational assessment has faced fundamental criticism regarding testing its research instruments. The quantitative instrument test that is usually carried out in CTT is the reliability index (Cronbach's alpha), which only measures the interaction between items and people how the quality of individual items can never be done because no measuring index can be done at the same time to detect inconsistent respondent answers not even available. In contrast to classical test theory, in Rasch modeling, item analysis is carried out at the level of each item. In addition to items, Rasch modeling simultaneously tests persons (respondents), where consistent patterns of respondents' answers will be seen, who tend to agree (in the attitude instrument) or identify random answers. Tests for research instruments can also be carried out in dimensionality tests and bias detection from the items being tested. That can be done because Rasch modeling fulfills all objective measurement requirements.

The Rasch model is a modern valuation theory that can classify item and person counts in a distribution map [22]. This model is part of the grain response theory [23]. In the Rasch model, test takers with high ability should have a greater probability of correctly answering a question than other students. And vice versa, students have a smaller chance to correctly answer a question with a higher difficulty level [16]. In the Rasch model approach, in addition to paying attention to items, it also pays attention to aspects of responses and their correlations[24]. Compared to other methods, especially classical test theory, the advantage of Rasch modeling is the ability to predict missing data based on a systematic response pattern [25].

An evaluation instrument is said to be good if it can provide appropriate information regarding students' abilities on the competencies tested [26]. In 2016, Susongko conducted research on the test item validation model, namely Messick validity, which includes several aspects such as content, structural, substantive, consequential, and external aspects [27]. This validity was analyzed using the Rasch model supported by Winstep software. Other research related to test theory analysis was conducted by Kustriyono, who used qualitative analysis to determine the characteristics of the tests being tested in terms of material, construction, and language [28]. Winstep software is a computational tool for the Rasch model to analyze scores generated from test instruments to know Outfit MNSQ, Outfit ZSTD, Point Measure Correlation, Item reliability, and Alpha Cronbach [29]. The MNSQ outfit is useful for seeing the suitability of the data with the model used. The expected mean-square value is 1 (one). If the mean-square value at infit is greater than one, the variation of the instrument is more than predicted by the Rasch model. If the infit value is less than 1, then the variation in the instrument is less when compared to the predictions made by the Rasch model[30].

In this study, the researcher wanted to know the quality of the evaluation instrument used to determine students' high-level thinking skills in static fluid material using the Rasch model approach. This quality is measured based on several indicators: item fit with the Rasch model and item reliability. Therefore an evaluation instrument was designed and then determined which items fit and which did not fit with the Rasch model. In addition, with the help of Winstep software, the Cronbach alpha value will be determined to determine the reliability of the items.

## 2. RESEARCH METHOD

This study focused on the analysis of evaluation instruments using the Rasch model. The Rasch model is a modern assessment theory that can classify item and person calculations in a distribution map to determine the quality of an instrument. The sampling technique in this study was by using purposive sampling. Respondents in this study were 36 students of class XI MIPA 1 Senior high school 2 Jambi City. There are 20 questions on the evaluation instrument that are tested on students. The level of difficulty of the questions starts from easy, medium, and difficult with relatively the same comparison. The questions are in the form of two-tier multiple-choice, namely multiple-choice questions which are included with a choice of reasons. The instrument grid used can be seen in appendix 1.

The results of the evaluation in the form of scores were analyzed using the Ministeps software. From the output of the Ministeps software, several item parameters are obtained that fit the Rasch model. In addition, Cronbach's alpha value was obtained which was the result of the overall item reliability test. Meanwhile, Outfit MNSQ, Outfit ZSTD and item correlation values with questions as a whole indicate the limits of items that are declared fit with the model. That is, if 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 a total score ranging from 0.4 to 0.85 [16].

## 3. RESULTS AND DISCUSSION

Based on data analysis using Ministeps software, there are 20 items that fit the Rasch model. These results are presented in full in Table 1.

Table 1. The results of descriptive

| Value | Decription | |
|---|---|---|
| Log | Person | -0.2 |
| | Items | 0 |
| Reliabelity | Person Reliability | 0.75 |
| | Reliability Items | 0.85 |
| | Alpha Cronbach | 0.79 |
| Outfit MNSQ | Person | 0.94 |
| | Item | 0.94 |
| ZSTD outfits | Person | 0 |
| | Item | -0.1 |

Table 1 shows the logit value of the person or person measure of -0.2 and the item measure value of 0, which means that the item measure value is greater than the person measure. Meanwhile, item reliability is worth 0.85, person reliability is worth 0.75 and Cronbach's Alpha value is 0.79. Another quantity shown in table 1 is the Outfit Mean Squared (Outfit MNSQ) value of 0.94 in both the person and item columns. Furthermore, Outfit Z Standardized (Outfit ZSTD) values were 0 for persons and -0.1 for items.

The distribution of item items that are considered fit and misfit or not fit with the model can be seen in Figure 1. The following Figure 1 shows the results of the data obtained from the Rasch modeling. not fit Rasch models.

Item STATISTICS: MISFIT ORDER

| ENTRY NUMBER | TOTAL SCORE | TOTAL COUNT | MEASURE | MODEL S.E. | INFIT MNSQ | INFIT ZSTD | OUTFIT MNSQ | OUTFIT ZSTD | PTMEASUR-AL CORR. | PTMEASUR-AL EXP. | EXACT MATCH OBS% | EXACT MATCH EXP% | Item |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 17 | 17 | 36 | -.08 | .38 | 1.28 | 1.60 | 1.46 | 1.95 | A .25 | .48 | 61.1 | 72.1 | S17 |
| 19 | 20 | 36 | -.51 | .38 | 1.24 | 1.41 | 1.23 | 1.00 | B .31 | .48 | 52.8 | 71.0 | S19 |
| 12 | 19 | 36 | -.36 | .38 | 1.16 | .97 | 1.23 | 1.04 | C .35 | .48 | 69.4 | 71.2 | S12 |
| 16 | 21 | 36 | -.65 | .38 | 1.11 | .71 | 1.23 | .99 | D .36 | .47 | 72.2 | 71.2 | S16 |
| 10 | 10 | 36 | 1.01 | .42 | .91 | -.40 | 1.13 | .47 | E .47 | .45 | 88.9 | 76.4 | S10 |
| 11 | 26 | 36 | -1.43 | .41 | 1.13 | .72 | 1.00 | .13 | F .34 | .42 | 69.4 | 75.0 | S11 |
| 14 | 27 | 36 | -1.60 | .42 | 1.13 | .69 | 1.06 | .28 | G .31 | .40 | 72.2 | 76.6 | S14 |
| 8 | 10 | 36 | 1.01 | .42 | 1.09 | .51 | .90 | -.18 | H .41 | .45 | 72.2 | 76.4 | S8 |
| 7 | 14 | 36 | .37 | .39 | 1.07 | .45 | .96 | -.07 | I .45 | .48 | 75.0 | 73.2 | S7 |
| 18 | 28 | 36 | -1.78 | .44 | 1.06 | .32 | .97 | .10 | J .34 | .38 | 83.3 | 78.5 | S18 |
| 15 | 19 | 36 | -.36 | .38 | 1.03 | .21 | .93 | -.27 | j .48 | .48 | 63.9 | 71.2 | S15 |
| 9 | 7 | 36 | 1.58 | .46 | 1.01 | .12 | .72 | -.46 | i .43 | .40 | 75.0 | 81.3 | S9 |
| 4 | 17 | 36 | -.08 | .38 | .98 | -.06 | .97 | -.05 | h .50 | .48 | 77.8 | 72.1 | S4 |
| 2 | 5 | 36 | 2.06 | .52 | .92 | -.17 | .81 | -.10 | g .39 | .35 | 88.9 | 86.0 | S2 |
| 5 | 9 | 36 | 1.19 | .43 | .91 | -.37 | .71 | -.69 | f .52 | .43 | 80.6 | 77.7 | S5 |
| 20 | 12 | 36 | .68 | .40 | .91 | -.41 | .79 | -.70 | e .55 | .47 | 77.8 | 74.5 | S20 |
| 6 | 10 | 36 | 1.01 | .42 | .86 | -.67 | .76 | -.63 | d .56 | .45 | 83.3 | 76.4 | S6 |
| 1 | 27 | 36 | -1.60 | .42 | .82 | -.91 | .67 | -.78 | c .54 | .40 | 83.3 | 76.6 | S1 |
| 3 | 19 | 36 | -.36 | .38 | .71 | -1.90 | .63 | -1.90 | b .70 | .48 | 80.6 | 71.2 | S3 |
| 13 | 17 | 36 | -.08 | .38 | .69 | -2.06 | .63 | -1.92 | a .72 | .48 | 83.3 | 72.1 | S13 |
| MEAN | 16.7 | 36.0 | .00 | .41 | 1.00 | .0 | .94 | -.1 | | | 75.6 | 75.0 | |
| P.SD | 6.8 | .0 | 1.08 | .03 | .16 | .9 | .23 | .9 | | | 9.0 | 3.8 | |

Figure 1. Data on the distribution of questions on fit and misfit or not fit with the Rasch model

Furthermore, it can be seen the data on the results of students' abilities from the person measure results from the Wright map in Figure 2 below:

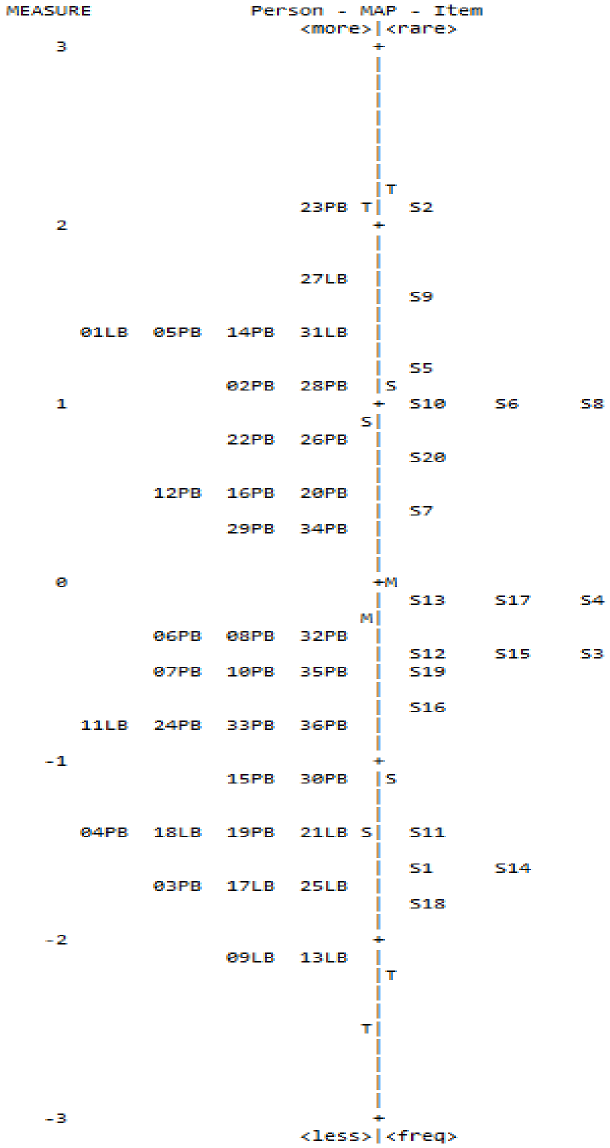


Figure 2. Person Measures

Next, the graphshows the measurement information obtained from student evaluation instruments on static fluid material can be seen in Figure 3 below:

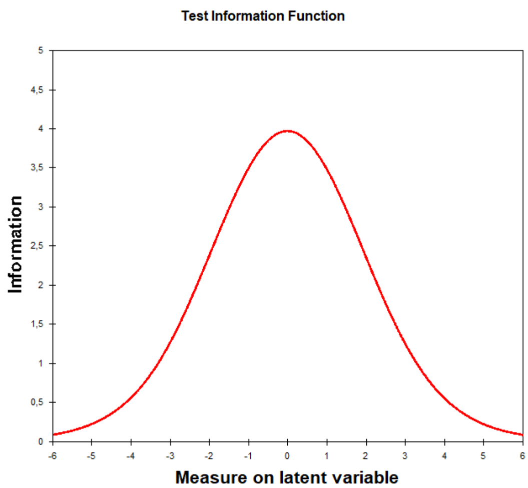


Figure 3. Test Information Function

We can discuss them individually based on data analysis using the Rasch model obtained from the above results. First, it can be seen from the reliability value of the instrument. A reliability test on a research instrument is used to determine whether the questionnaire used in collecting research data is reliable or not [31]. If a variable shows a Cronbach Alpha value > 0.60, it can be concluded that this variable can be said to be reliable or consistent in measuring [31]. From the values obtained in the Rasch model, it can be stated that the level of consistency of students' answers is quite high, and the quality of the items in the evaluation instrument used has good reliability. This is supported by Azizah's research [32], which also uses the Rasch model, which shows the reliability value of the instrument expressed in Cronbach's alpha is 0.85, which means a high-reliability value.

Furthermore, the fit criteria for the items can be seen based on the results of the MNSQ outfit values, both person and item, which are in the ideal range of 0.5<MNSQ<1.5 [16]. From the results obtained, the MNSQ outfit values are already in the ideal range. This means that the evaluation instrument used is under the Rasch model. Then, based on the results of the Outfit Z Standardized (Outfit ZSTD) values obtained, it is between the range -2.0 <ZSTD < 2.0, which can be interpreted as having a rational possible value [32]. This means the items follow the Rasch model and can be used as an evaluation instrument for static fluid material. This is also supported by Azizah's research [32], which obtained the results of 25 questions according to the Rasch model with an average ± average Outfit MNSQ for a person and item ± 0.98 and 0.98 respectively, while the Outfit ZSTD value for person and item is ± 0 each and -0.01.

Based on Figure 1 above regarding the distribution of item items that are considered fit and misfit or not fit with the model can be analyzed with the terms and conditions of the value limit. Item boundaries are declared fit with the model if one or both conditions are met. The first requirement is that the Outfit MNSQ value lies between 0.5 to 1.5; the Outfit ZSTD value lies between -2.0 to 2.0, and the item correlation value with the total score (point measure correlation) lies between 0.4 to 0.85[21]. Based on the analysis of the evaluation instrument using the Ministeps program in Figure 1, it was found that all the items met the fit criteria, meaning that out of 20 items, there were no misfit items.

Furthermore, the value of students' ability level in working on the questions is shown from the Ministeps output, namely the Wright map in Figure 2 above. Data obtained from students with code 23PB have the highest ability or ability, while students with codes 09LB and 13LB have the lowest ability. Figure 3 shows the measurement information obtained from student evaluation instruments on static fluid material. The x-axis shows students' ability level in working on a given test, while the y-axis shows the value of the information function. Based on the graph, the information obtained by the measurement is very high at the medium ability level.

## 4. CONCLUSION

Based on the results of the Rasch modeling evaluation instrument used for physics subjects on static fluid material fit with the Rasch model. This is indicated by the item score (item reliability) of 0.85, person reliability of 0.75, and Cronbach's alpha value of 0.79, while the Outfit Mean Square Statistics (Outfitt MNSQ) value is 0.94 in the person and item column. The value of Outfit Z Standard (Outfit ZSTD) is 0 for the person table and -0.1 for the item table. While the number of items that fit as many as 20 items. The instrument meets the reliability criteria, which is quite high, so that means the instrument is feasible to use. For future researchers, implementing the Rasch model trials on other instruments is better. To know whether the instrument used is fit and reliable.

## ACKNOWLEDGEMENTS

The author would like to thank several parties who have helped in completing this research. Supervising lecturers who have helped give directions, as well as students and teachers who are willing to conduct research in their classes.

## REFERENCES

[1] I. Magdalena, H. N. Fauzi, and R. Putri, “Pentingnya Evaluasi Dalam Pembelajaran Dan Akibat Memanipulasinya,” *J. Pendidik. dan Sains*, vol. 2, no. 2, pp. 244–257, 2020.

[2] A. N. Phafiandita, A. Permadani, A. S. Pradani, and M. I. Wahyudi, “Urgensi Evaluasi Pembelajaran di Kelas,” *JIRA J. Inov. dan Ris. Akad.*, vol. 3, no. 2, pp. 111–121, 2022, doi: 10.47387/jira.v3i2.262.

[3] S. Aqliyah, “Pembuatan Media Evaluasi Pembelajaran dengan Quizizz di SMK Annur Pasaleman pada Mata Pelajaran Simulasi dan Komunikasi Digital,” vol. 1, no. 2, pp. 1–29, 2022.

[4] I. P. Suardipa and K. H. Primayana, “Peran Desain Evaluasi Pembelajaran Untuk Meningkatkan Kualitas Pembelajaran,” *Widyacarya*, vol. 4, no. 2, pp. 88–100, 2020.

[5] D. Desilva, I. Sakti, and R. Medriati, “Pengembangan Instrumen Penilaian Hasil Belajar Fisika Berorientasi HOTS (Higher Order Thinking Skills) pada Materi Elastisitas dan Hukum Hooke,” *J. Kumparan Fis.*, vol. 3, no. 1, pp. 41–50,

Apr. 2020, doi: 10.33369/jkf.3.1.41-50.

[6] N. W. Putri and R. Dwijayanti, "Pengembangan Alat Evaluasi Bantuan Aplikasi 'Quizizz'Pada Mata Pelajaran Marketing Kelas X Jurusan Bdp Di Smk Negeri 10 Surabaya," *J. Pendidik. Tata Niaga*, vol. 8, no. 3, pp. 985–991, 2020.

[7] L. . Anderson and D. . Krathwohl, *A Taxonomy for Learning, Teaching, and Assesing*. New York: Longman, 2001.

[8] E. Cahyono, S. Lathif, Y. Pantiwati, and A. S. Budi, *Pengembangan Perangkat Pembelajaran Berorientasi HOTS Tingkat Sekolah Dasar*. 2020.

[9] S. Arikunto, *Dasar-dasar Evaluasi Pendidikan*. Jakarta: Bumi Aksara, 2013.

[10] D. F. Treagust, "Diagnostic assessment in science as a means to improving teaching, learning and retention," 2006.

[11] T. M. Haladyna and S. M. Downing, "Applied Measurement in Education A Taxonomy of Multiple- Choice Item-Writing Rules," *Appl. Meas. Educ.*, vol. 2, no. 1, pp. 37–50, 1989.

[12] M. Nofiana, S. Sajidan, and P. Puguh, "Pengembangan Instrumen Evaluasi Two-Tier Multiple Choice Question untuk Mengukur Keterampilan Berpikir Tingkat Tinggi pada Materi Kingdom Plantae," *J. Inkuiri*, vol. 3, no. 2, pp. 60–74, 2014.

[13] Y. Pantiwati and N. Mahmudati, "Increased ability to arrange two tier multiple choice questions," *J. Community Serv. Empower.*, vol. 2, no. 1, pp. 22–27, 2021, doi: 10.22219/jcse.v2i1.15855.

[14] L. W. Olsen, *Essays on Georg Rasch and his contributions to statistics*. Unpublished Ph.D thesis at Institute of Economics University of Copenhagen., 2003.

[15] J. M. Linacre, "A User's guide to WINSTEPS Ministeps; Rasch-model Computer Program. Program Manual 3.73." 2011.

[16] B. Sumintono and W. Widhiarso, *Penilaian Pendidikan dan Ujian*. 2015.

[17] J. M. Linacre, "Investigating rating scale category utility," *J. Outcome Meas.*, vol. 3, no. 2, pp. 103–122, 1999.

[18] G. Englehard, *Invariant Measurement, using rasch models in the social, behavioral and health sciences*. New York: Routledge, 2013.

[19] Y. Irhasyuarna., F. Fahmi., and I. Setiadi, "Mengembangkan keterampilan berpikir tingkat tinggi melalui pembelajaran IPA," *Prosiding Magister Pendidikan Ilmu Pengetahuan Alam*, vol. 1, no. 1, 2021.

[20] B. Mok, M. and Wright, *Overview of Rasch Model Families.In Introduction to Rasch Measurement: Theory, Models and Applications*. Minnesota: Jam Press, 2004.

[21] B. Sumintono and W. Widhiarso, "Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial," 2014.

[22] M. Rozeha, AR, Azami, Z. & Mohd Saidfudin, "Application of Rasch Measurement in Evaluation of Learning Outcomes: A Case Study in Electrical Engineering. Regional Conference on Engineering Mathematics, Mechanics, Manufacturing & Architecture 2007 (EM3ARC)." 2007.

[23] Thissen, D., Nelson, L., Rosa, K., *Item Response Theory for Items Scored in More than Two Categories in D. Thissen & H. Wainer, Test Scoring.* New Jersey: Lawrence Erlbaum Associates Publishers, 2001.

[24] D. Ardiyanti, "Application of the Rasch Model on the Development of the Self-Efficacy Scale in Student Career Decision Making," *J. Psychol.*, vol. 43, no. 3, pp. 248–263, 2016.

[25] R. Aziz, "Application of the Rasch model in testing mental health measures at work. Psychoislamika," *J. Islam. Psychol. Psychol.*, vol. 12, no. 2, pp. 29–39, 2015.

[26] Y. Wisman, E. Effrata, and T. Tutesa, "Penerapan Konsep Instrumen Evaluasi Hasil Belajar," *J. Ilm. Kanderang Tingang*, vol. 12, no. 1, pp. 1–9, 2021, doi: 10.37304/jikt.v12i1.105.

[27] E. Suseno, "Measuring Test Validity. educative speaker." 2021.

[28] Kustriyono, "Penyusunan Perangkat Soal Ujian Akhir Mata Pelajaran SainsBiologi Dalam Rangka Pengembangan Bank Soal," *J. Penelit. dan Eval. Pendidik.*, vol. 2, no. VI, pp. 175–198, 2004.

[29] A. Gunasti, A. Sanosra, N. A. Mufarida, and E. B. Satoto, "Pemanfaatan Rasch Model untuk Mengukur Kemampuan," vol. 7, no. 2, pp. 1544–1557, 2023.

[30] B. Sumintono, "Penilaian Keterampilan Berpikir Tingkat Tinggi: Aplikasi Pemodelan Rasch pada Asesmen Pendidikan," in *Prosiding Magister Pendidikan Ilmu Pengetahuan Alam*, 2021.

[31] E. Rosita, W. Hidayat, and W. Yuliani, "Uji Validitas Dan Reliabilitas Kuesioner Perilaku Prososial," *FOKUS (Kajian Bimbing. Konseling dalam Pendidikan)*, vol. 4, no. 4, p. 279, 2021, doi: 10.22460/fokus.v4i4.7413.

[32] Azizah and S. Wahyuningsih, "Penggunaan Model Rasch Untuk Analisis Instrumen the Use of Rasch Model for Analyzing Test," *J U P I T E K J. Pendidik. Mat.*, vol. 3, no. 1, pp. 45–50, 2020.