Development of Chemistry Learning Outcomes Assessment Instruments in Chemistry Matter Even Semester XI Grade of Senior High School

1st Mhd Burhan Anggara[1], 2nd Ajat Sudrajat2, 3rd Muhammad Yusuf[3]

{ burhananggara38@gmail.com[1], ajat.sudrajat_49@yahoo.com[2], yusuf.6423@gmail.com[3]}

[1,2,3]Postgraduate Chemistry Education, Universitas Negeri Medan

Abstract. Assessment is the process of assessing the learning outcomes of students to assess which of the learning goals have been accomplished. The focus of this study is of assessing the test items' validity , reliability , and the level of difficulty. During the 2022–2023 academic year, the data collection strategy involved using an answer sheet for the class XI final exam held in the even semester at SMAS Al-Amjad Medan. There were 151 respondents overall, and the question consisted of 40 multiple-choice questions (dichotomies) with five possible categories of answers. The fit items are obtained by using the Rasch model. Version 3.7.3 of the Winsteps software was used to conduct this investigation. The following conclusions can be made based on the outcomes of the item analysis performed using the Rasch model: The questions are 95% valid and 5% invalid in terms of validity; Because 1 item (3%) come in under the most difficult category, 2 things (7.5%) go into the tough category, 33 items (82%) fall into the moderate category, and 3 items (7,5%) fall into the easy category , the distribution of the difficulty level of the itemsu fits with the rules for developing questions. The good category's Cronbach's alpha value is 0.75, the enough category's Person Reliability value is 0.71, and the very good category's Item Reliability rating is 0.95; The parameters for creating questions determined how difficult the things were distributed, with 1 item (3%) falling into the extremely difficult group, 2 items (about 7.5%) in the difficult category, 33 items (approximately 82%) in the moderate category, and 3 items (approximately 7.5%) in the easy category. Subsequently, 30 out of 40 items were classified as suitable for use as a tool to assess students' HOTS based on the findings of the item analysis conducted with Rasch. The result of this research is a legitimate chemistry learning outcome instrument that can be used online to assess students' HOTS proficiency.

Keywords: validity, reliability, difficulty level, rasch model

AISTEEL 2023, September 19, Medan, Indonesia

DOI 10.4108/eai.19-9-2023.2340440

1 Introduction

One method for assessing student learning outcomes and determining whether or not learning objectives are met is to conduct assessment activities. An evaluation is conducted to see whether an activity has been carried out in accordance with the stated objectives. [1]. Evaluation of learning outcomes is a sequence of actions used by teachers to obtain and analyze information about their students' learning processes and outcomes during the learning process [2]. A well-conducted assessment will yield favorable outcomes for the caliber of education. Therefore, in order to ascertain their pupils' academic ability, teachers must complete the required evaluation activities. [3It is important for educators to possess not only the ability to assess the learning process but also the capability to create instruments or tools for evaluation that are suitable for the desired learning results. [4]. In order to determine whether or not the learning objectives have been fulfilled, it is also necessary to adapt the methodologies, tools, and assessment instruments to the means of measurement and data gathering.

During the learning process, one of the methods used by the teacher to determine the ability and academic achievement of students is to carry out assessments. The only factor that teachers consider when determining students' ability levels is the number of questions they answer correctly [1]. Summative assessment is a form of evaluation that is often carried out in educational institutions. The semester end exam is an example of a summative assessment. The purpose of summative assessment is to ascertain the level of student achievement after a certain period of time has passed. The instrument referred to here is the measuring instrument used to carry out the evaluation. Multiple choice tests are the most frequently used type of assessment tool today. If the learning processes and evaluation tools are of high quality, good learning outcomes can be achieved [5] . According to [6] for a test to be considered valid as a tool for evaluating something, the test must meet the criteria of validity, reliability, objectivity, usability, and affordability. A competent instrument will ensure that the findings it produces are valid and reliable to provide an accurate assessment of student abilities . When evaluating the calibre of assessment tools, educators frequently employ the classical theory analysis approach. However, there are numerous flaws in analysis that is based on classical theory. Therefore, instrument analysis utilizing analysis based on the Rasch model is one way to get around the shortcomings of classical theory. [5],

The use of the Rasch analysis model to analyze the quality of the instrument can be said to be feasible. The analyzes performed using the Rasch model are not only very simple to perform but also produce accurate analytical findings. According to [7] that the Rasch model provides a number of benefits. The Rasch model can detect wrong answers and missing data scores, as well as the fact that ability does not completely depend on the correct answers [8]. The evaluation tool must show proof of the items' validity, reliability, degree of difficulty, and distinctive power in order to assess the items' quality. [9] and distractors [10]. Research conducted by [1] shows that the collaborative problem solving assessment instrument meets the valid criteria with an average score of 28 out of 32 and an estimated reliability of 0.70 if using reliable criteria. A total of 117 students out of a possible total of 124 provided profiles of their collaborative problem solving abilities, with results indicating the moderate category. Furthermore, [11] demonstrates five expert panels evaluating quality item 23 out of 40 (57.5%), statistical compatibility aspect, expert Val.A and the most consistent Val.F (based on

MNSQ and ZSTD), and item reliability demonstrated a good level (0.97). Research on expert judgment analysis of item items has been conducted using the Rasch measurement model approach.

This explanation states that instructors are advised to utilize the Rasch model, a sort of assessment analysis, while measuring and analyzing student learning outcomes to ascertain students' initial skills. Based on these factors, the study's goal was to examine the instrument item quality for the odd semester final exams in the high school chemistry class XI subject by using the Rasch model.

2 Methods

Research design

At the end of the even semester of class XI, instrument questions were used to collect primary data for this descriptive study that used quantitative data analysis. In this instance, quantitative descriptive research entails seeing, analyzing, and statistically characterizing the research topic in order to make inferences about it from the events that are observed. [12]

Research Objectives

This research was carried out at SMAS Al-Amjad Medan in the 2022/2023 school year. The subjects of this study were 151 students of class XI MIPA.

Research data

The data for this study came from the odd semester end exam, which included 40 questions and was composed entirely of multiple-choice (dichotomous) questions with five categories of possible responses. There were 151 participants in the study overall as students. The Rasch model was used to determine feasible questions based on an analysis of each student's responses.

Data Collection Instruments

The study tool utilized was a question item consisting of forty items that were dichotomous multiple-choice questions with five possible answer categories. Answer sheets from the even semester final exam for the class XI chemistry final at SMA Al-Amjad Medan during the 2022–2023 academic year served as documentation for the data gathering technique.

Data analysis

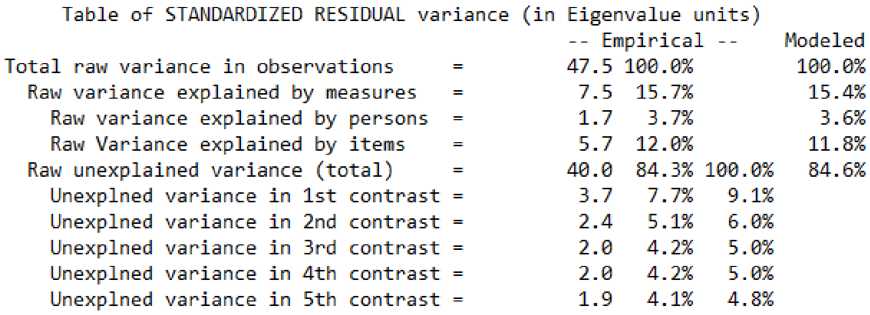
In data processing, research data analysis techniques include quantitative and descriptive data analysis. Data on student learning outcomes were collected in the form of chemistry subject end-of-semester exam questions, student response sheets, and key answer sheets using a documentation approach. The result sheet documentation provides a quantitative analysis of the questions, with a particular focus on validity, reliability and level of difficulty. By using Winstep software version 3.7.3, the quality of the questions was analyzed using a computer program (software) designed to process data from multiple choice tests.

Several parameter items that match the Rasch model have been extracted from the Winsteps program's output. The value of the Outfit MNSQ, Outfit ZSTD, and Pt Measure Corr items can be used to determine whether the items meet the requirements. Additionally, the results of the item reliability test questions are the values acquired, which are the Person Reliability values and Item Reliability values. Valid or acceptable items must satisfy at least two of the three requirements; if they don't, they should be eliminated. If they do, they should be corrected [1]. The item measure, or item difficulty level, gives a complete view of each item's logit value. Information about the items sorted from those with the highest logit measure values to those with the lowest logit measure values is included in the output results.

3 Results And Discussion

3.1 Validity test

Rasch modeling connects the test-takers' ability to the item parameters. The Winstep application was used to analyze the Rasch model. Before using the Rasch model for analysis, a few presumptions must be met, including the requirements for unidimensionality and local independence. The output item dimensionality displays the results of the Winstep program's unidimensionality test. Figure 1 displays the analysis's output results.



```
     Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)
                                            -- Empirical --    Modeled
Total raw variance in observations     =        47.5 100.0%        100.0%
  Raw variance explained by measures   =         7.5  15.7%         15.4%
    Raw variance explained by persons  =         1.7   3.7%          3.6%
    Raw Variance explained by items    =         5.7  12.0%         11.8%
  Raw unexplained variance (total)     =        40.0  84.3% 100.0%   84.6%
    Unexplned variance in 1st contrast =         3.7   7.7%   9.1%
    Unexplned variance in 2nd contrast =         2.4   5.1%   6.0%
    Unexplned variance in 3rd contrast =         2.0   4.2%   5.0%
    Unexplned variance in 4th contrast =         2.0   4.2%   5.0%
    Unexplned variance in 5th contrast =         1.9   4.1%   4.8%
```

Fig. 1. The results of unidimensionality and local independence assumptions

The actual value was 15.7%, as opposed to the 15.4% anticipated by the Rasch model, according to the findings of the construct validity study on the Raw variance explained by measures explained by empirical measurements, which are displayed in Figure 1. In this instance, the value that the Rasch model predicts and the validation of the empirical construct are nearly identical. If the Raw variance explained by measurements was more than 40%, the construct validation evaluation was considered extremely good. This study's examination of construct validity was deemed inadequate. However, all of the unexplained variance values that were found were 15%, which is considered poor. [1]. Because the chemistry teacher in this study did not conduct construct validation, construct validation was rated as inadequate. professors do not look to other chemistry professors for approval; instead, they develop their own questions.

Furthermore, whether or not the item questions satisfy a number of rules used by Boone et al. with the aim to determine the validity of the question items using the Rasch model chooses the degree of conformity of the item questions. (Susdelina et al., 2018), specifically:

The values of the following outfits were received:

a) MNSQ (Mean Square): 0.5 < Outfit – MNSQ < 1.5;

b) ZSTD (Z–Standard): –2.0 < ZSTD < +2.0;

c) Pt Measure Correlation (Point Measure Correlation): 0.4 < Pt Measure Corr < 0.85

The outfit means square (MNSQ), outfit z standard (ZSTD), and point measure correlation (PT MEASURE CORR) values are the criteria used to determine the amount of item fit, according to Boone et al. (2014). The findings of the MNSQ, ZSTD, and Pt Measure Corr outfits in Table 1 show the validity test results.

Table 1. Results of Outfit MNSQ, ZSTD, and Pt Measure Corr

Items	Outfits		PtMeasure Corr.	Information
	MNSQ	ZSTD		
1	0.91	-1,4	0.41	*fit*
2	1.04	0.5	0.27	*fit*
3	1.19	2,2	0.9	*Misfit*
4	0.94	-1.0	0.38	*Fit*
5	1.19	1.0	0.17	*Fit*
6	1.26	1,2	0.08	*Fit*
7	1.05	0.5	0.21	*Fit*
8	1.22	1,6	0.17	*Fit*
9	0.92	-0.3	0.40	*Fit*
10	1.21	0.8	0.18	*fit*
11	1,12	1,6	0.12	*fit*
12	0.81	-1.9	0.52	*fit*
13	0.81	-1.7	0.52	*fit*
14	0.84	-2.0	0.48	*fit*
15	1.09	0.7	0.23	*fit*
16	1.11	1,6	0.17	*fit*
17	1.02	0.2	0.29	*fit*
18	0.98	-0,2	0.33	*fit*
19	1.02	0.4	0.26	*fit*
20	0.86	-1.5	0.48	*fit*
21	0.88	-1.0	0.45	*fit*
22	0.99	-0.1	0.31	*fit*
23	1,13	0.7	0.14	*fit*
24	0.96	-0,2	0.33	*fit*
25	0.93	-0.6	0.40	*fit*
26	1.17	1.0	0.17	*fit*
27	1.07	0.5	0.23	*fit*
28	0.98	-0.1	0.30	*fit*
29	0.91	-1,4	0.41	*fit*
30	1.06	1.0	0.22	*fit*
31	0.91	-0.9	0.44	*fit*
32	0.97	0	0.25	*fit*
33	0.95	-0,2	0.34	*fit*

Items	Outfits		PtMeasure Corr.	Information
	MNSQ	ZSTD		
34	0.94	-0.7	0.38	*fit*
35	1,13	0.9	0.18	*fit*
36	0.96	-0.4	0.31	*fit*
37	0.88	-1.5	0.45	*fit*
38	1,6	2,2	-0.07	*Misfit*
39	0.82	-0.7	0.42	*fit*
40	0.89	-1,4	0.44	*fit*

Based on Table 1, when viewed from the three criteria, there are 2 item items (13.33%) that are misfit, namely item numbers 3 and 38 which do not meet the MNSQ, ZSTD and Pt Measure Corr outfit requirements. The existence of outliers in the analysis causes the results of the suitability analysis of the items to be disturbed so that they become less reliable, therefore item numbers 7, 11, 17, 22, 24, and 38 need to be removed. Whereas for the other 39 item items (86.67%) that only met one or two criteria were declared fit for the model [13], the item items only needed to be corrected [1].

3.2 Reliability Test

Determining an instrument's reliability involves determining how consistently it measures when it is used on a regular basis. The Cronbach Alpha rating indicates the overall reliability value. The Person Reliability results can be used to ascertain the respondents' degree of consistency. In the meantime, the item reliability value is used to determine the quality per item in the instrument. If an item satisfies the requirements listed in Table 2, its quality from the reliability side of the item employs the Cronbach Alpha value. While the Rasch model's criteria for evaluating an item's dependability have an impact on person reliability as well, Table 3 illustrates these criteria in more detail [14]).

Table 2. Reliability criteria in Rasch modeling

Correlation coefficient	Correlation	Interpretation
$0.90 \leq r \leq 1.00$	Very high	Very good
$0.70 \leq r \leq 0.90$	Tall	Good
$0.40 \leq r \leq 0.70$	Currently	Pretty good
$0.20 \leq r \leq 0.40$	Low	Bad
$r < 0.20$	Very low	Very bad

Table 3. Reliability criteria in Rasch modeling

Reliability Value (Person/Item)	Interpretation
> 0.94	Special
0.91 – 0.94	Very good
0.81 – 0.90	Good
0.67 – 0.80	Enough

< 0.67	Weak

The findings of the Rasch model reliability study on even semester final exam questions for chemistry courses are shown in Table 4.

Table 4. The results of the reliability analysis

Reliability Analysis	Reliability	Category
Cronbach Alpha	0.75	Good
Person Reliability	0.71	Enough
Reliability Items	0.91	Very well

The reliability score of 0.91 of the item puts it in the very excellent group, whereas the person reliability score of 0.71 placed it in the sufficient category, based on the results shown in Table 4., and the Cronbach alpha value of 0.75 puts it in the good group. As a result, it can be concluded that the instrument's component dependability is high.

3.3 Difficulty Level of Question Items

The item difficulty level determines how many respondents tend to answer the item correctly. In Rasch modeling, to see how difficult an item is based on the results of the Item Measure table. The log values for each item are shown in Item Measure, which are ordered from highest to lowest. The logit value is a measure of item difficulty. The more difficult the item is, the higher the logit value. If a question's degree of difficulty is proportionate, or balanced, then it is considered good. [4]. Four groups are identified by the distribution of the item's difficulty level (Erfan et al., 2020). Table 5 displays the findings of the distribution of the items' difficulty levels.

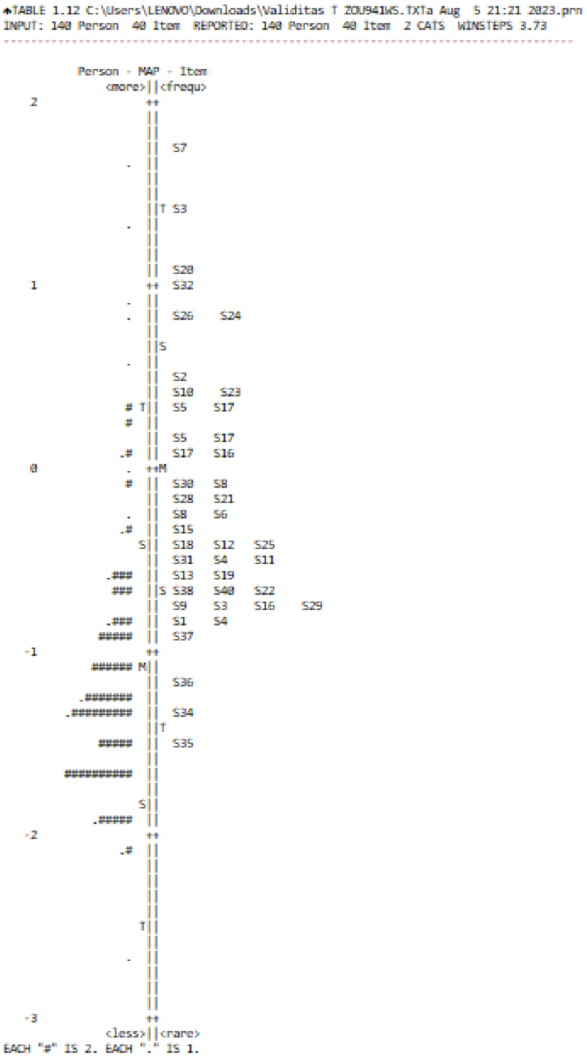
Table 5. The results of the distribution of the difficulty level of the questions

Degree of difficulty	Measured Value	Question Items	Number of Question Items	Percentage
Very difficult	*Measure logs*>0.82	7	1	3%
Difficult	0.00 ≤ Measure logit ≤ 0.82	20, 32, 33	3	7.5%
Currently	–0.82 ≤ Measure logit ≤ 0.00	1, 2, 3, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 37, 38, 39, 40	33	82%
Easy	*Measure logs*< –0.82	34, 35, 36	3	7.5%

Items are examined using the Rasch model, where each item has a Measure logit value that indicates the degree of difficulty. Table 5 demonstrates that there are 1 item (3%) in the category of extremely tough objects, 3 item items (7.5%) in the category of difficult goods, 33 item items (82%) in the category of moderate items, and 3 item items (7, 5%) in the category of very simple items.

The number of students that successfully answered the item defined the item difficulty level category. The Wright map's distribution map of person-item associations can also be used to view logit values. Combining the logit values for item and person measures results in the map shown in the table. The Wright map displays the logit item value on the right side and the logit person value on the left.

Figure 2 illustrates that there is no discernible variation in the pupils' aptitude levels when it comes to answering questions. Their location on the map is rather near to other pupils' positions, which is indicative of this. The item's logit value is explained by the Wright map on the right. This figure reveals that item S10 is a question with a very challenging level of difficulty. As a result, there is little chance that pupils will correctly answer this question. [15]. On the other hand, item S4 is a question with a very simple difficulty level. This is further demonstrated by the percentage of students who answered item S4 correctly (137 out of 222 total students did so). if the logit values of the item questions are the same.



```
+TABLE 1.12 C:\Users\LENOVO\Downloads\Validitas T ZOU941WS.TXTa Aug  5 21:21 2023.prn
INPUT: 140 Person  40 Item  REPORTED: 140 Person  40 Item  2 CATS  WINSTEPS 3.73
------------------------------------------------------------------------------------

           Person - MAP - Item
               <more>||<frequ>
    2                ++
                     ||
                     ||
                     ||  S7
                  .  ||
                     ||
                     ||
                     ||T S3
                  .  ||
                     ||
                     ||
                     ||  S20
    1                ++  S32
                  .  ||
                  .  ||  S26    S24
                     ||
                     ||S
                  .  ||
                     ||  S2
                     ||  S10    S23
                  # T||  S5    S17
                  #  ||
                     ||  S5    S17
                 .#  ||  S17   S16
    0             .  ++M
                  #  ||  S30   S8
                     ||  S28   S21
                  .  ||  S8    S6
                 .#  ||  S15
                    S||  S18   S12   S25
                     ||  S31   S4    S11
               .###  ||  S13   S19
                ###  ||S S38   S40   S22
                     ||  S9    S3    S16    S29
               .###  ||  S1    S4
              #####  ||  S37
   -1                ++
             ###### M||
                     ||  S36
           .#######  ||
         .#########  ||  S34
                     ||T
              #####  ||  S35
                     ||
         ##########  ||
                     ||
                    S||
             .#####  ||
   -2                ++
                 .#  ||
                     ||
                     ||
                     ||
                     ||
                    T||
                     ||
                  .  ||
                     ||
                     ||
                     ||
   -3                ++
               <less>||<rare>
EACH "#" IS 2. EACH "." IS 1.
```

Figure 2. Distribution of student abilities and item difficulty questions (person-item map)

4 Conclusion

The following conclusions can be made in light of the item item analysis results utilizing the Rasch model analysis: The questions are 95% valid and 5% invalid in terms of validity; their Cronbach alpha value is 0.75, placing them in the good category; their Person Reliability value is 0.71, placing them in the sufficient category; and their Item Reliability value is 0.91, placing them in the very good category; according to the guidelines for creating questions, the distribution of the item items' difficulty level is as follows: 1 item item (3%) is in the very difficult category, 33 item items (82%) fall into the medium group, 3 item questions (7.5%)

fall into the easy category, and 3 item items (7.5%) fall into the difficult category. Thirty-three out of the forty questions in the Rasch item analysis were deemed suitable for use as a measure of students' HOTS.

References

[1] L. Purniasari, M. Masykuri, and S. R. D. Ariani , "Analisis Butir Soal Ujian Sekolah Mata Pelajaran Kimia SMA N 1 Kutowinangun Tahun Pelajaran 2019/2020 Menggunakan Model Iteman Dan Rasch," J. Pendidik. Kim., vol. 10, no. 2, pp. 205–214, 2021, [Online]. Available: https://jurnal.uns.ac.id/jpkim

[2] A. N. Syadiah and G. Hamdu, "Analisis rasch untuk soal tes berpikir kritis pada pembelajaran STEM di sekolah dasar," Prem. Educ. J. Pendidik. Dasar dan Pembelajaran, vol. 10, no. 2, p. 138, 2020, doi: 10.25273/pe.v10i2.6524.

[3] Susdelina, S. A. Perdana, and Febrian, "Analisis Kualitas Instrumen Pengukuran Pemahaman Konsep Persamaan Kuadrat Melalui Teori Tes Klasik dan Rasch Model," no. 1, pp. 41–48, 2018.

[4] R. Rusiyah, S. S. Eraku, and S. Supadmi, "Analisis Soal Ujian Akhir Semester Mata Pelajaran Geografi Dengan Menggunakan Pemodelan Rasch ," J. SWARNABHUMI J. Geogr. dan Pembelajaran Geogr., vol. 5, no. 1, p. 11, 2020, doi: 10.31851/swarnabhumi.v5i1.4136.

[5] F. Alfarisa and E. A. Wibawa, "Sub Theme : Education an Analysis Item of Final Examination of Economic Subject of Senior High School in Yogyakarta City," pp. 483–494, 2016.

[6] S. Arikunto, Prosedur Penelitian Suatu Pendekatan Praktek, VI. Jakarta: PT.Rineka Cipta, 2006.

[7] A. Emda, "KEDUDUKAN MOTIVASI BELAJAR SISWA DALAM PEMBELAJARAN," vol. 5, no. 2, 2015.

[8] W. Eliza and E. Yusmaita, "Pengembangan Butir Soal Literasi Kimia pada Materi Sistem Koloid Kelas XI IPA SMA/MA," J. Eksakta Pendidik., vol. 5, no. 2, pp. 197–204, 2021, doi: 10.24036/jep/vol5-iss2/621.

[9] A. L. Son, "Instrumentasi Kemampuan Pemecahan Masalah Matematis: Analisis Reliabilitas, Validitas, Tingkat Kesukaran Dan Daya Beda Butir Soal.," vol. 10, no. 1, pp. 41–52, 2019.

[10]R. Siahaan, M. Sitorus, and S. Silaban, "The development of teaching materials oriented to critical thinking skills for chemistry class XI high school," J. Pendidik. Kim., vol. 13, no. 1, pp. 60–68, 2021, doi: 10.24114/jpkim.v13i1.24145.

[11] S. Yamtinah* et al., "Content Validity in Android-Based Augmented Reality Media for High School Science Students on Covalent Bonds Topic: Rasch Model Analysis," J. Pendidik. Sains Indones., vol. 10, no. 2, pp. 240–249, 2022, doi: 10.24815/jpsi.v10i2.23280.

[12] E. A. Putra, "Anak Berkesulitan Belajar di Sekolah Dasar Se-Kelurahan Kalumbu Padang," J. Ilm. Pendidik. Khusus, vol. 1, no. 3, pp. 71–76, 2015, [Online]. Available: http://103.216.87.80/index.php/jupekhu/article/viewFile/6065/4707

[13] 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.

[14] M. Erfan, M. A. Maulyda, V. R. Hidayati, F. P. Astria, and T. Ratu, "Tes Klasik Dan Model Rasch," Indones. J. Educ. Res. Rev., vol. 3, no. 1, pp. 11–19, 2020, [Online]. Available: https://ejournal.undiksha.ac.id/index.php/IJERR/article/view/24080/pdf

[15] Kurniawan and K. D. K. Andriyani, "Analisis Soal Pilihan Ganda dengan Rasch Model," J. Stat., vol. 6, no. 1, pp. 34–39, 2018.