Article

An Analysis of the Online Final Examination Items for Ninth-graders in the Mathematics Course Using the Rasch Measurement Model

Tommy Tanu Wijaya [1]*

[1] Beijing Normal University, China;202139130001@mail.bnu.edu.cn

* Correspondence: 202139130001@mail.bnu.edu.cn; Tel.: +86 18577395150

Abstract: Indonesia's National Examination has been abolished since 2020. Hence, the Indonesian junior high schools make their final examination items for the 9th-grade, and from the results, the school determines students' graduation. Therefore, this study is aimed to evaluate the Online Final Examination items in one of the public junior high schools in Bandung, Indonesia. The sample was 234 students in grade 9 using their mathematics examination tests, comprising 20 multiple-choice items with 4 options, while the data processing used Winsteps software with the Rasch modeling technique. Subsequently, the Rasch model results showed an acceptable person separation statistic of 1.54 and sufficient person reliability at 0.74. The item separation statistics were in a good category at 4.59, while the reliability at 0.95 was excellent. Although four online final examination items were in the fit category, 16 were good and capable of dividing students according to their abilities. The result also provided very detailed data about the quality of the items and the ability of each grade 9 student. Since each test item is included in the fit category, this study contributes information on teachers' preparing and evaluating the Online Final Examination.

Keywords: coronavirus Pandemic, Rasch Model, School exams.

1. Introduction

Indonesia is one of the countries in Southeast Asia with a high number of Covid-19 cases, and as of May 18, 2020, there were 1.7 million cases and 48,305 deaths. This has caused many challenges and problems in education, as students have not attended school since May 2020, and teaching and learning activities are taking place online [1]. Several problems arise when implementing online learning, as the teachers do not have good pedagogical and technological knowledge and technological knowledge to teach on these platforms [2]. Also, students are not ready to take online lessons [1].

A national examination is a measuring tool for sustainable national growth in the education sector, which aims to see if students can think critically, be innovative, and provide solutions for the world. Meanwhile, National examinations are usually held in many countries to determine students' graduation and future when entering high school or university [3], [4]. Similar to other countries, Indonesia student’s examination is held every year to determine the students' graduation and analyze their quality in each region. Due to Covid-19, the Indonesian Ministry of Education decided to abolish the national examinations from elementary to high school level. Therefore, each school was allowed to hold a final examination and determine the students' graduation independently. This development can be beneficial depending on the angle from which experts and teachers analyzed it. Consequently, the government cannot determine the quality of education in each region anymore because every school has different graduation examinations and standards. The worst aspect is that schools may assess a child's graduation subjectively and pass a child with a poor achievement. Since this problem is yet to be solved, analyzing and researching the quality of final examination items for grade 9 junior high school students is an important to be performed.

The challenges of final graduation examination are even greater because they are held online, using Google Classroom, Zoom, Google Meeting, and other applications. Furthermore, the examination is in the multiple-choice and description form, inputted into the Google platform to allow students to work online. In the end, the teacher does not have to check the multiple-choice answers, as students can immediately see the final results. However, Indonesia does not have a stable internet network and an effective platform to ensure students do not cheat during the examination. They may use calculators while solving the mathematics questions or cellphones to ask friends for answers, and even people around can help them get perfect scores. Hence, the analysis of these results is very important to evaluate the students' abilities and ensure cheating does not occur.

During the Covid-19 pandemic, Indonesia implemented a 75% work-from-home rule, which prevented proper communication and meetings for producing and evaluating the final examination questions. The mathematics teachers experienced difficulties asking for help from other teachers to conduct the construct validation analysis while making their test. Furthermore, this was the first time the school was to determine the junior high school students' graduation based on the final examination scores because of the Covid-19 pandemic. Therefore, the school lacked experience making adequate items to determine the junior high school student's graduation.

This research focuses on analyzing the online mathematics examination to replace the national examinations in determining student graduation. The novelty is the analysis of two important points, the items' quality and students' abilities using Rasch Analysis with WINSTEPS software. Furthermore the research was performed at a public school in Bandung, Indonesia, using the Rasch measurement model to evaluate the difficulty, reliability, and quality of items, alongside analyze the probability curve and students' abilities. Consequently, the results are expected to contribute to education as an evaluation tool when the national examination is abolished. They can also serve as a reminder for teachers when preparing items for final examinations, such as midterms, semester finals, or the online mathematics finals.

By processing the final examination data using the Rasch measurement model, the results can answer the following questions:

1. How are the validity values in the online mathematics final examination with aspect of person fit, item fit, and unidimensionalty?

2. How are the reliability values of online mathematics final examination items based on i) KR-20 Cronbach’s coefficient alpha; and ii) Person and item reliability indices?

3. How are the students' ability levels and item difficulty levels of online mathematics final examination items?

4. Are there any indications of cheating when students work on the online mathematics final examination?

2. Research Theory

2.1 National Examination

Every year, many countries use the national examination to determine the graduation of grade 9 students in junior high schools. These countries have their designations for this examination, for instance, it is called Scholastic Aptitude Tests (SAT) or American College Testing (ACT) in America [5]. In China, it is known as the Chinese National College Entrance Examination (CEE; gaokao) [3] and in Indonesia, it is called the Ujian Nasional (UN) [6].

The national examinations are conducted once yearly around May and are important for most students to determine their future and further education. In Indonesia, students who fail can retake it in the following year. The national examination is used to evaluate students, schools, and provinces to help the government know the quality of education. From the results, the government can focus on areas that require special attention and improvement to ensure education in the country is fair and equitable.

Several studies on national examinations were previously performed, such as Virginia LoCastro, compared the national examination on the English subject in Japan using sociocultural analysis [7]. In another study, Bai et al. (2014) examined the National College Entrance Examination and analyzed students' abilities at two leading universities in China, with implications for university admission policies and practices. This study suggested that admissions should consider the National College Entrance Examination results, alongside the high school achievement over the past three years, count the prizes and awards received in high school, and consider homeroom teacher recommendations. Although the National College Entrance Examination score is just a number, this study perceives that many factors should be reconsidered, for example, general and language skills and attitudes.

Sanz and Pavón (2015) explained that the national foreign language examination system in Spain uses an online system. The illustration of the platform, management tools, security, and user interaction are well explained, and the study concluded that the use of online systems for national examinations existed before the coronavirus pandemic. Meanwhile, no specific research discussing the item analysis of the national examination using the online system was discovered. Therefore, this research contributes to discussing the national examination item analysis using an online system during Covid-19. The analysis can also be used as a reference source, benchmark, and comparison material to assess the quality of national examination items in Indonesia during Covid-19. Consequently, a public school was used as the research sample.

2.2 The Rasch Model

The Rasch model is a mathematical model and measurement tool [10]. It focuses on the approach to construct measurement in the social sciences field, which usually uses the more familiar WINSTEPS software and can show the response structure of the assessment [11]–[13]. Moreover, it can enter the class, student work, questionnaire items, or final examinations and provide specific measurement data [14], [15]. Although the Rasch model shows the measurement criteria and test fitting responses, the analysis of insufficiently fit data should be continued by carefully evaluating the reliability and validity [16]. The model is also a method for analyzing examination results to investigate the correlation between item difficulty and students' abilities [17], [18].

Meanwhile, Karlin and Karlin (2018) stated the importance of validating the test items. They affirmed that the validity test can determine the accuracy of test items in measuring students' abilities, for instance, in mathematics. Validation can also ascertain whether the difficulty level of the test items is based on the student's abilities. Can the test items separate the students into three or four levels based on their abilities? Are the test items self-explanatory? Do test items confuse students? The research concluded that ensuring test items given are appropriate and good in measuring students' abilities is important.

Subsequently, the number of items with high, medium, and low difficulty should be planned properly, hence the student's ability level can be measured specifically. The test items that are too difficult will achieve mostly incorrect answers, while students will correctly answer too easy questions. These conditions cannot measure which students have low, medium, or very good abilities. Therefore, test items should be created with 50% medium items, 30% easy, and 20 percent with high difficulty.

The item analysis using the Rasch model has been performed in many previous studies in various fields such as medicine, pharmacy, physics, chemistry, social science, and so on to validate and evaluate the quality of the questionnaire items [20]–[23]. However, only a few studies have used the Rasch model to analyze items in the mathematics field, while none that analyzed final school examination items using the model was discovered.

A study on the use of the Rasch model to validate and analyze items in 2011 was found. Mohsen Tavakol and Reg Dennick used the model to improve assessment in medical education by analyzing 355 medical students using 24 final clinical knowledge items [24]. The analysis results concluded that Rasch analysis supports the diagnosis of quality, alongside provides feedback for each test item and students' ability to inform lecturers on methods to improve the quality of examination items. In 2011 again, Abdullah et al. (2012) analyzed examination items on Microelectronic material using the Rasch model and concluded on its ability to analyze students' abilities when answering examination items.

Also, Jennings and team in 2016, analyzed multiple-choice examination items conducted on 101 students at the Arizona University and found that they were not very difficult, hence the classification of students' abilities was not measured properly [25]. They advised that the item-making should focus on measuring the quality performance rather than just looking at student grades and rankings. Furthermore, research by Nopiah et al. (2011) on Engineering Mathematics Courses on the code paper KKKQ2114 used the Rasch model to validate and analyze items' quality and found that 10% were in the misfit category Therefore, the items should be corrected or removed, as the bad items cannot measure students' abilities according to the teacher's wishes.

Based on these previous studies, this research concluded that the Rasch model is very important and has many benefits for evaluating and improving the quality of items as well as measuring and analyzing students' abilities. Furthermore, the results from the model can be used as evaluation material to modify the teaching method or the item formed to fit the guidelines for good test items.

2. Research Methodology

This study used the Rasch model to analyze test items from the Online Final Examination data. The measurement model can show summary statistics such as mean, Standard Deviation, maximum value, amount of data, etc., and classify students' abilities as low, medium, or high by evaluating separation and logit. It can also determine whether the items for obtaining a measurement were met, the level of difficulty and whether students are careful while answering or making guesses.

2.1 Research Population

The research data comprised mathematics examination items for the 9th-grade graduation class held in May 2021. Meanwhile, the sample was the 9th-grade students in one of the public junior high schools in Bandung, Indonesia. The purposive sampling technique was used in the selection, and the general information about the study sample is shown in Table 1. This public school was chosen because it has good national accreditation and quality.

Table 1. General Information of study sample.

Category	Public School
Accreditation	A
Number of students	930 students
Status	National standard school
Lowest National Examination Score (2015)	270/400
Location	Bandung, West Java, Indonesia

The final examination items were made by the 9th-grade mathematics teachers and validated by 2 curriculum experts, then signed by the principal and used to determine the students' graduation in mathematics. Subsequently, the research sample was 234 students from 6 classes (9A, 9B, 9C, 9D, 9E, 9F), comprising 155 female and 79 male students, which made up 66.24% and 33.76%, respectively.

2.2 Data collection

A teacher with the initials IM made 20 graduation examination items for mathematics in the multiple-choice form, consisting of 6 items on numbers, algebra, and geometry each, and 2 items on probability and statistics each. Table 2 shows more specific information.

Table 2. The final examination material according to Indonesian national standards.

No.	Tested Competence	Scope of topic	Topic	Cognitive level
1	The students can understand and are knowledgeable about integer operations	Numbers	Integer operations	Knowledge and understanding
2	The students can understand and are knowledgeable about quadratic operations	Numbers	Quadratic operation	Knowledge and understanding
3	The students can apply their knowledge of fractional numbers	Numbers	Fractional number	Application
4	The students can apply their knowledge of comparisons	Numbers	Comparison	Application
5	The students can apply their knowledge of social arithmetic	Numbers	Social arithmetic	Application
6	The students can apply their knowledge of number sequences and series	Numbers	Sequences and series of numbers	Reasoning
7	The students can understand and are knowledgeable about the linear inequality of one variable	Algebra	Linear inequality of one variable	Knowledge and understanding
8	The students can understand and are knowledgeable about sets of numbers	Algebra	Set of numbers	Knowledge and understanding
9	The students can use reasoning related to straight-line equations	Algebra	Quadratic function	Knowledge and understanding
10	The students can apply algebraic forms	Algebra	Algebra forms	Application
11	The students can apply their knowledge of algebraic forms	Algebra	Quadratic function	Application
12	The students can apply their knowledge of relations or functions	Algebra	Function value	Application
13	The students can apply their knowledge of geometry and measurement	Geometry and measurement	Pythagorean Theorem	Application
14	The students can understand and are knowledgeable about lines and angles	Geometry and measurement	Lines and angles	Knowledge and understanding
15	The students can understand and are knowledgeable about triangles' similarity and congruence	Geometry and measurement	Triangles similarities and congruence	Knowledge and understanding
16	The students can apply their knowledge of curved side spaces	Geometry and measurement	Curved side spaces	Application

17	The students can apply their knowledge of circles	Geometry and measurement	Circles	Application
18	The students can understand transformation knowledge	Geometry and measurement	Transformation	Knowledge and understanding
19	The students can understand the data presentation of frequency table forms	Statistics and probability	Data centering measure	Knowledge and understanding
20	The students can understand the probability of events	Statistics and probability	Probability of events	Application

The Final Examination Items were designed for use with ordinal data that are scored in two categories (0 or 1) namely dichotomous data. Each test item was assigned a 5-point score, and the student's total scores will range from 5 to 100. Based on the results, this study used the Rasch measurement model to classify the students' achievement when working on the Online Final Examination items.

2.3 Process

This research analyzed the teaching and learning process by interviewing 2 mathematics teachers at the school. According to the interview results, these activities were still occurring online, and the school used Google Meeting and Google Classroom, and occasionally Zoom to interact with students. Meanwhile, homework and exercises were given through Google Classroom. From the interview results, the teachers stated that no math software or videos were used for mathematics lessons during the Covid-19 pandemic.

The 9th-grade final examination information was obtained through Google Meeting and lasted for 90 minutes, with the students using laptops to work on the items and a handphone placed behind them for supervision. However, not all the students have these gadgets, and some only used handphones to take the final examination. Another problem was the slow and disconnected internet, which did not allow the teacher to monitor the students properly. Finally, the instructions were given to students to log in at 9 P.M and finish the test items at 10.30 P.M.

2.4 Statistical analysis

The Online Final Examination Items was analyzed using WINSTEPS version 3.73. This research used the Person fit statistics to evaluate the consistency of the students' answers in the online examination. An inconsistent Person fit is caused by cheating or guessing while answering the items. Table 3 shows the criteria for assessing the items and the students' ability levels.

Table 3. *The reference table for the items' validity and students' abilities* [27].

statistics	Criteria	Additional information
Point measure Correlation (PTMEA-CORR)	0.4-0.85	To evaluate the difficulty level of the items from the hardest to the easiest
Model S.E	X<0.5	X<0.5 means that it can adequately determine the students' abilities
Outfit Mean Square Values (MNSQ)	0.5 <X<1.5	A too-large MNSQ value (>1.5) means that students with a high ability answered incorrectly on an 'easy' item. A too-small MNSQ value (< 0.5) means that students with a low ability answered correctly a 'difficult' item but incorrectly for the rest of items
Outfit Z-standartized Values (ZSTD)	-2.0 <ZSTD<+2.0	A too-high Outfit ZSTD value (> 2.0) indicates that a student with a high ability answered incorrectly on an 'easy' item. A too-small Outfit ZSTD value (<-2.0) indicates that a student with a low ability answered correctly a 'difficult' item but incorrectly for the rest of items.

Meanwhile, the reliability of the Online Final Examination items can be analyzed using the Rasch model based on the KR-20 analysis, item and person reliability and item and person separation. The references table for the interpretation of data reliability can be seen in Table 4.

Table 4. *The reference table for reliability analysis* [28].

statistics	Fit Indicator	Additional information
Cronbach's alpha (KR-20)	<0.5	Low
	0.5 – 0-6	Moderate
	0.6 – 0.7	Good
	0.7 – 0.8	High
	> 0.8	Very Good
Item and Person reliability	<0.67	Low
	0.67 – 0.80	Sufficient
	0.81 – 0.90	Good
	0.91 – 0.94	Very Good
	> 0.94	Excellent
Item and person separation	1 – 2	Enough
	>2	Good

Rasch analysis is able to determine the students' ability levels and difficulty of online final mathematics examination items. Wright map was used to visualize students' ability and difficulty levels on the same logit scale. On the wright map, item difficulty is on the right and students' ability levels are on the left. The easiest questions are at the bottom and the hardest questions are at the top of the wright map and for low ability students are at the bottom and high students are at the top.

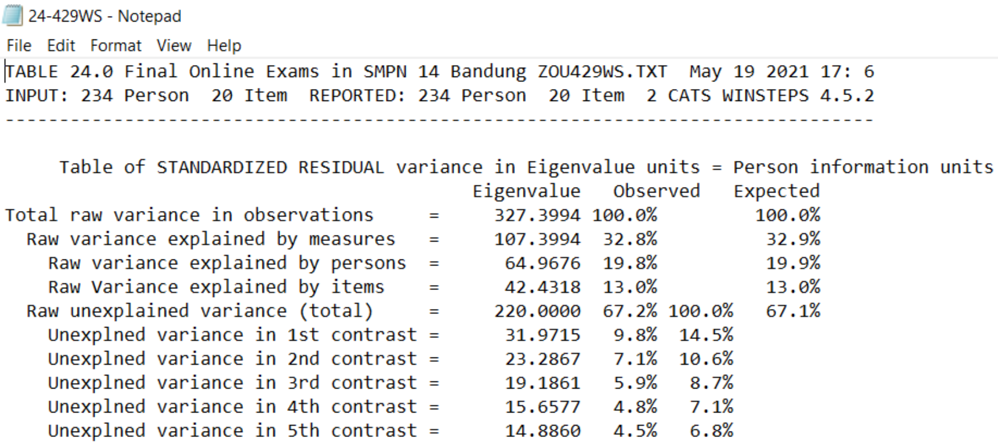
Lastly, Guttman Scalogram was used to assess students' responses to each item. The careless, guessing and cheating students when students work on the online mathematics final examination can be analysed from the guttman scalogram.

3. Results

3.1. Construct Validity

3.1.1. Unidimensionality

Unidimensionality is frequently analyzed using the Principal Component Analysis of Rasch Residual (PCAR). The analysis of unidimensionality was performed with the Rasch model, and the results obtained were organized in the output table as shown in Figure 1. The construct validity results of the *Raw variance explained* by empirical measures produced a score of 32.8%, while the Rasch model predicted 32.9%. Since the empirical construct validation is almost the same as the value predicted by the Rasch model, the minimum unidimensionality requirement of 20% was met (> 40% means good, and > 60% means excellent). Meanwhile, the unexplained variance for the 1st until 5th contrast is less than 10%, which falls in the ideal range value of less than 15% (Sumintono & Widhiarso, 2015). This findings indicate that the online mathematics final examination has a strong evidence of unidimensionality, that is, the items undoubtedly measured the construct of 9th-grade mathematics.



```
TABLE 24.0 Final Online Exams in SMPN 14 Bandung ZOU429WS.TXT  May 19 2021 17: 6
INPUT: 234 Person  20 Item  REPORTED: 234 Person  20 Item  2 CATS WINSTEPS 4.5.2
--------------------------------------------------------------------------------

     Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = Person information units
                                           Eigenvalue   Observed   Expected
Total raw variance in observations     =     327.3994 100.0%         100.0%
  Raw variance explained by measures   =     107.3994  32.8%          32.9%
    Raw variance explained by persons  =      64.9676  19.8%          19.9%
    Raw Variance explained by items    =      42.4318  13.0%          13.0%
  Raw unexplained variance (total)     =     220.0000  67.2% 100.0%   67.1%
    Unexplned variance in 1st contrast =      31.9715   9.8%  14.5%
    Unexplned variance in 2nd contrast =      23.2867   7.1%  10.6%
    Unexplned variance in 3rd contrast =      19.1861   5.9%   8.7%
    Unexplned variance in 4th contrast =      15.6577   4.8%   7.1%
    Unexplned variance in 5th contrast =      14.8860   4.5%   6.8%
```

Figure 1. The output table of unidimensionality analysis.

3.1.2. Item Fit

Construct validation can be performed by involving several validators to obtain better results. Similar items are first tested in small groups, then the results are re-evaluated per item. The teachers can also ensure that the sentences per item are unambiguous and easy to understand. However, although the construct validation was not performed, it can be predicted by the Rasch model. This model was very effectively applied, as it can obtain reliable validity analysis results and is easy to use because it can be directly analyzed by computer applications. Table 5 shows the output table of item fit, which represents the content validity analysis results that can be seen from the level of items' suitability.

Table 5 shows the results of the item fit analysis. There were two item statistics of misfit orders, namely the infit and outfit. The outfit statistics were used more frequently due to their higher sensitivity to data with extreme scores. After comparing the 20 items, 16 items were found to be fit, while 4 items were not because they did not meet the 3 criteria above (table 3). Table 5 shows that the topmost item, number 15, did not fit, as it does not meet the requirements for Outfit ZSTD based on the criteria in table 3 (value 3.32) and had a low measure correlation (0.36). This means that item 15 should be investigated more because it does not contribute adequately to classifying students' mathematical abilities. Another interpretation is that it is too difficult for students, and cannot classify those with high and low mathematical abilities. The next step was to revise the test item. Conversely, the infit mean square statistic showed that the average of each item was 1.0, meaning it was within acceptable limits. The MNSQ outfits for test items 19 and

doi:10.20944/preprints202109.0510.v1

7 were outside the acceptable limit, while number 1 had a low measure correlation, meaning it was too easy or difficult for students. Therefore, the final decision on the test items should be investigated and revised. However, according Sumintono and Widhiarso (2015) if the item meets one of the criteria (Outfit MNSQ, Outfit ZSTD, or PTMEA-CORR), the item should be retained.

Table 5. Item fit analysis in the Online Mathematics Final Examination.

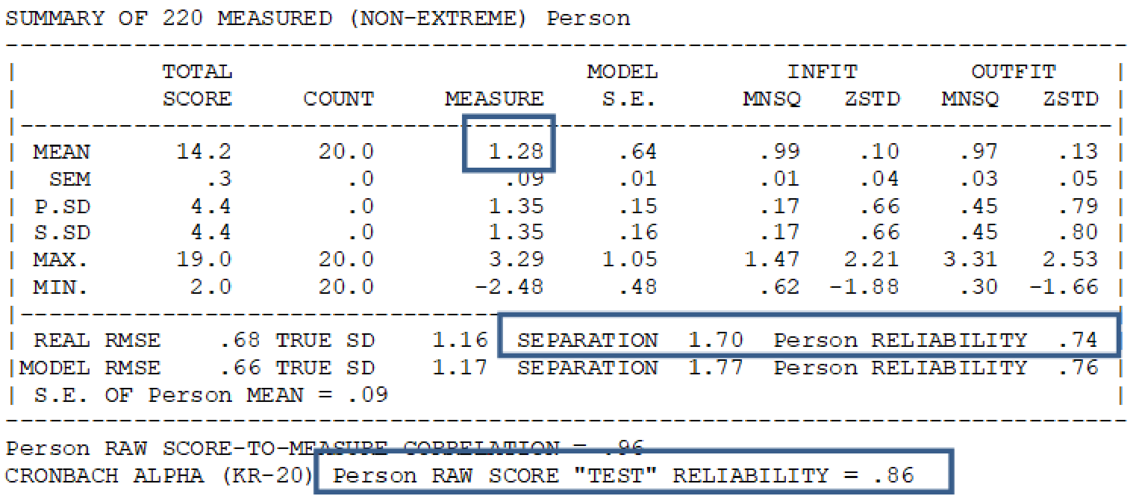
Item Number	Measure	Outfit MNSQ	Outfit ZSTD	PTMEA-CORR
15	1.68	1.49	**3.32**	**0.36**
19	0.91	1.34	**2.77**	0.44
5	-0.72	1.15	0.69	0.43
10	1.04	1.15	1.31	0.47
17	0.37	1.15	1.09	0.53
11	0.28	1.07	0.52	0.50
12	0.81	1.07	0.66	0.53
14	1.68	1.05	0.43	0.53
1	-1.91	0.77	- 0.40	**0.37**
6	-0.16	1.02	0.15	0.51
3	-0.89	0.91	-0.27	0.49
8	0.04	0.83	-1.06	0.56
9	-0.41	0.82	-0.89	0.54
18	-0.34	0.89	-0.54	0.54
4	-1.03	0.81	-0.64	0.49
16	-0.30	0.89	-0.53	0.55
20	-0.30	0.85	-0.75	0.55
13	-0.03	0.81	-1.16	0.58
7	0.13	0.68	**-2.32**	0.63
2	0.85	0.72	-1.14	0.56

3.2 Reliability

3.2.1 Person Reliability

The Online Mathematics Final Examination had 20 items divided into four subtopics, namely Numbers, Algebra, Geometry and measurement, Probability, and Statistics. Subsequently, the Rasch model analyzed the correlation between the students' mathematical abilities and the test items, using an examination taken by 234 students. The Rasch statistical analysis model was divided into two categories, which are the 234 measured persons and the 20 measured items. Tables 5 and 6 discuss the summary statistics for each category in detail.

Figure 2. Summary statistics of person reliability.

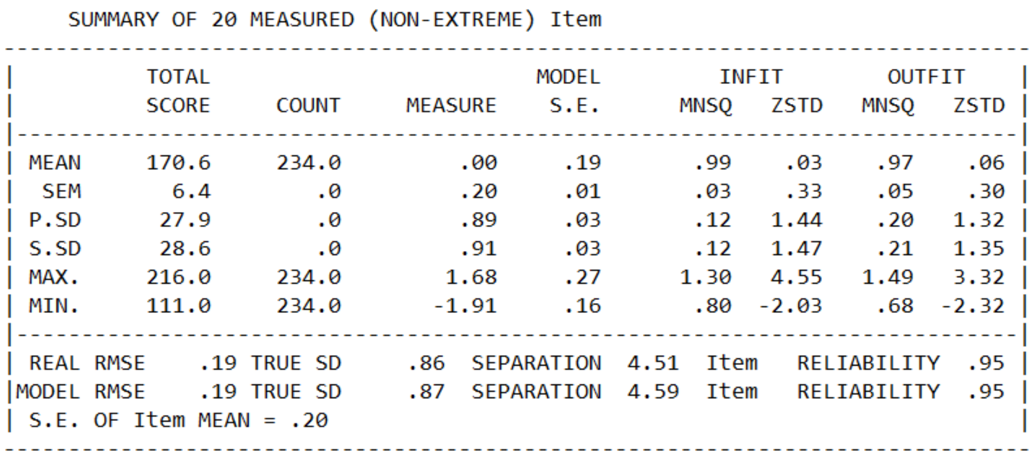


```
SUMMARY OF 220 MEASURED (NON-EXTREME) Person
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE   S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      14.2       20.0       1.28     .64       .99    .10    .97    .13 |
|  SEM        .3         .0        .09     .01       .01    .04    .03    .05 |
| P.SD       4.4         .0       1.35     .15       .17    .66    .45    .79 |
| S.SD       4.4         .0       1.35     .16       .17    .66    .45    .80 |
| MAX.      19.0       20.0       3.29    1.05      1.47   2.21   3.31   2.53 |
| MIN.       2.0       20.0      -2.48     .48       .62  -1.88    .30  -1.66 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .68 TRUE SD    1.16  SEPARATION  1.70  Person RELIABILITY  .74 |
|MODEL RMSE    .66 TRUE SD    1.17  SEPARATION  1.77  Person RELIABILITY  .76 |
| S.E. OF Person MEAN = .09                                                   |
-------------------------------------------------------------------------------
Person RAW SCORE-TO-MEASURE CORRELATION = .86
CRONBACH ALPHA (KR-20) Person RAW SCORE "TEST" RELIABILITY = .86
```

figure 2 is a statistical summary of the 234 students who took the Online Mathematics Final Examination. Cronbach's alpha, which is the person raw score reliability that measures the reliability using the interaction between the person and the item as a whole, was 0.86. According to the rating scale of the items' quality criteria, the Cronbach's alpha fell in the range of 0.81 to 0.90, meaning the overall quality of the items was in the very good category. The person measure was +1.28 logit, showing the average value of all students working on the item given. Meanwhile, the average value was larger than the logit value of 0.0, indicating a tendency for students' mathematical abilities to be higher than the level of difficulty. A value of 0.74 was obtained for the person reliability, indicating the consistency of the students' answers was in the sufficient or acceptable category. The teacher could have more items that were more difficult that would mark the trait a little better. That one reason why the person reliability is not higher than 0.9 [29]. The INFIT MNSQ and OUTFIT MNSQ produced average values of 0.99 and 0.97, respectively, where the ideal value is 1, i.e., the closer to 1, the better. In addition, the INFIT ZSTD and OUTFIT ZSTD gave average table person values of 0.10 and 0.13, where, in this case, the ideal value is 0.0, that is, a value closer to 0.0 depicts better quality. The value of person separation was 1.7. According to Wright and Masters (2002), the number of person separate strata (*H*) can be culculated from the separation index by using the equation: *H* = [(4 x separation index) + 1] / 3. The person separation index of 1.7 produces a strata (*H*) of 2.6 (can be rounded up to 3). This value indicates the students can be well distinguished into three different abilities that is, high, medium, and low ability.

Figure 3 shows summary statistics for the 20 items of the Online Mathematics Final Examination. The statistical data results were aimed at analyzing the item categories, either difficult, medium, or easy, and provide overall quality information of the student response patterns, the instrument used, and the interaction between person and the items. Subsequently, the average INFIT MNSQ and OUTFIT MNSQ values were 0.99 and 0.97, where the ideal value is 1, that is, the closer to 1, the better. Conversely, the average INFIT ZSTD and OUTFIT ZSTD values of the item table were 0.03 and 0.06, and in this case, the ideal value was 0.0, where a closer score to 0.0 signifies better quality. Based on the data analysis results, the difficulty levels were obtained from the *item measure output* results. Then, the items were grouped by combining the logit mean and the standard deviation value, which produced averages of 0.00 and 0.91, respectively. These values were used to identify the item separation. The item separation was 4.59, indicating the items can be divided into 5 categories, namely very easy, easy, medium, difficult, and very difficult. Hence, a larger separation value signifies a better quality of the items for dividing the students' mathematical abilities. According to Linacre (2012), an item separation value which is more than 2.00 is interpreted as good. In this study, the value for item reliability was 0.95. Sumintono and Widhiarso (2015) stated that an item reliability which is higher than 0.94 is interpreted as 'excellent'.

Figure 3. Summary statistics for item reliability.



```
     SUMMARY OF 20 MEASURED (NON-EXTREME) Item
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE   S.E.       MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN     170.6      234.0         .00    .19        .99    .03    .97    .06 |
|  SEM       6.4         .0         .20    .01        .03    .33    .05    .30 |
| P.SD      27.9         .0         .89    .03        .12   1.44    .20   1.32 |
| S.SD      28.6         .0         .91    .03        .12   1.47    .21   1.35 |
| MAX.     216.0      234.0        1.68    .27       1.30   4.55   1.49   3.32 |
| MIN.     111.0      234.0       -1.91    .16        .80  -2.03    .68  -2.32 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .19 TRUE SD     .86  SEPARATION  4.51  Item   RELIABILITY  .95 |
|MODEL RMSE    .19 TRUE SD     .87  SEPARATION  4.59  Item   RELIABILITY  .95 |
| S.E. OF Item MEAN = .20                                                     |
-------------------------------------------------------------------------------
```

3.3 The Wright Maps (or Person-item distribution map)

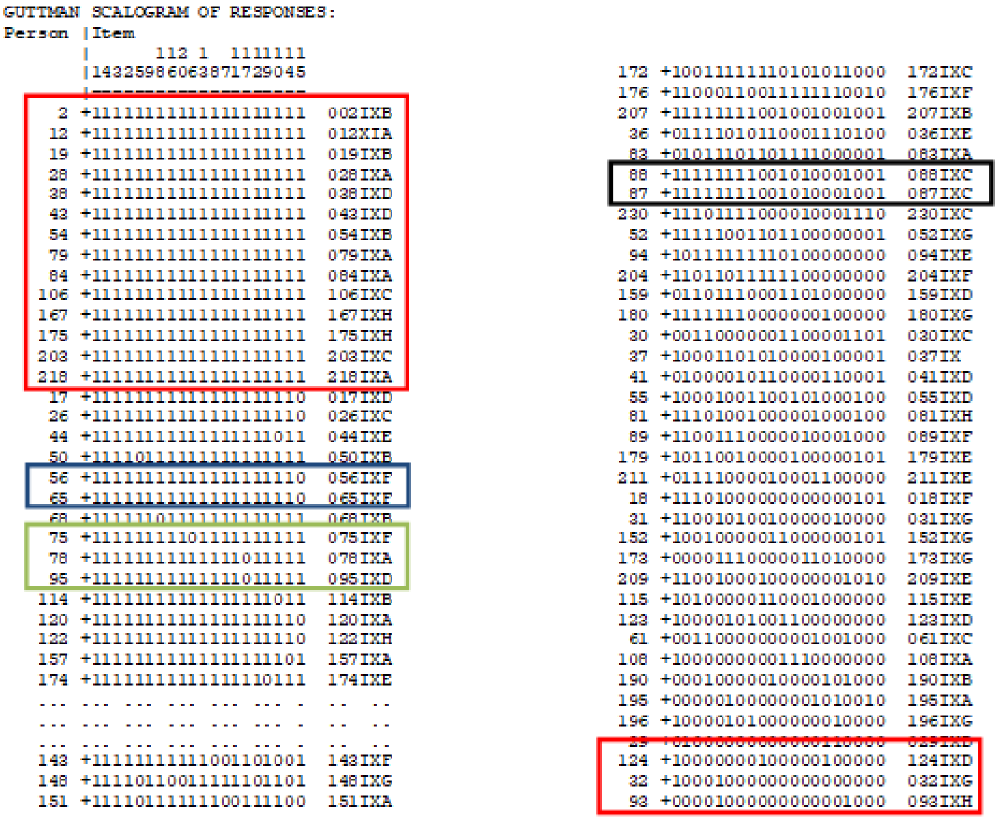
Table 6 shows the specific classification of the students' abilities, which were more in the moderate, high and low ability groups than the very low and very high ability groups.

Table 6. Classification of the students' mathematical abilities.

Code of classification	Code of students	Total of student	interpretation
+T	002IX 012XI 019IX 028IX 038IX 043IX 054IX 079IX 084IX106IX 167IX 175IX 203IX 218IX	14	Very high ability student
+S	010IX 015IX 020IX 034IX 046IX 048IX 091IX 109IX 117IX 118IX 121IX 125IX 126IX 128IX 131IX 132IX 136IX 141IX 144IX 145IX 149IX 156IX 161IX 162IX 171IX 183IX 184IX 185IX 191IX 193IX 197IX 198IX 201IX 215IX 219IX 222IX 226IX	37	High ability student
M	004IX 013IX 014IX 022IX 045IX 053IX 058IX 059IX 060IX 064IX 067IX 071IX 072IX 074IX 077IX 080IX 090IX 098IX 099IX 100IX 103IX 105IX 110IX 116IX 119IX 127IX 134IX 137IX 147IX 153IX 155IX 160IX 164IX 170IX 188IX 205IX 208IX 210IX 212IX 217IX 221IX 224IX 225IX 231IX 006IX 024IX 033IX 042IX 047IX 070IX 096IX 102IX 104IX 113IX 140IX 146IX 150IX 154IX 166IX 187IX 199IX 216IX 227IX 001IX 005IX 009IX 011IX 021IX 051IX 057IX 069IX 097IX 111IX 139IX 142IX 143IX 148IX 151IX 169IX 178IX 194IX202IX 206IX 234IX 016IX 027IX 035IX 062IX 073IX 087IX 130IX 135IX 163IX214IX 228IX 229IX 049IX 092IX 133IX 165IX 168IX 177IX 186IX 189IX 200IX 008IX 101IX 158IX 172IX 176IX 207IX 036IX 083IX 088IX 192IX 230IX 052IX 094IX 204IX	138	moderate ability student
-S	003IX 007IX 025IX 066IX 076IX 085IX 086IX 107IX 181IX	35	Low ability

	182IX 232IX 233IX 023IX 039IX 040IX 082IX 129IX 138IX 159IX 180IX 030IX 037IX 041IX 055IX 081IX 089IX 179IX 211IX 018IX 031IX 152IX 173IX 209IX 115IX 123IX		Student
-T	061IX 108IX 190IX 195IX 196IX 029IX 124IX 032IX 093IX 112IX	10	Very low ability student

The Guttman scalogram observed the student answers on each item and ranked their abilities from the highest to the lowest [12]. As shown in Figure 4, five people in class 9A, three in 9B as well as two in 9C, 9D, and 9H achieved perfect scores. These 14 students answered all the items on the Online Mathematics Final Examination correctly, while the students with codes 032IXG, 093IXH, and 112IXE had the lowest abilities. The three lowest students were at the bottom and could only answer two items out of the 20 correctly. The text continues here.



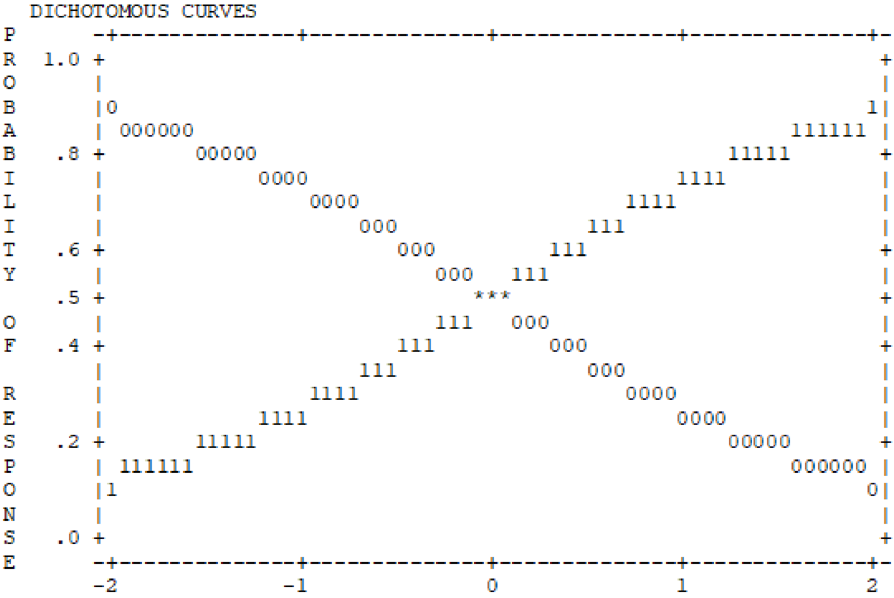
```
GUTTMAN SCALOGRAM OF RESPONSES:
Person |Item
       |        112 1  1111111
       |14325986063871729045
       |--------------------
     2 +11111111111111111111  002IXB
    12 +11111111111111111111  012XIA
    19 +11111111111111111111  019IXB
    28 +11111111111111111111  028IXA
    38 +11111111111111111111  038IXD
    43 +11111111111111111111  043IXD
    54 +11111111111111111111  054IXB
    79 +11111111111111111111  079IXA
    84 +11111111111111111111  084IXA
   106 +11111111111111111111  106IXC
   167 +11111111111111111111  167IXH
   175 +11111111111111111111  175IXH
   203 +11111111111111111111  203IXC
   218 +11111111111111111111  218IXA
    17 +11111111111111111110  017IXD
    26 +11111111111111111110  026IXC
    44 +11111111111111111011  044IXE
    50 +11110111111111111111  050IXB
    56 +11111111111111111110  056IXF
    65 +11111111111111111110  065IXF
    68 +11111101111111111111  068IXB
    75 +11111111101111111111  075IXF
    78 +11111111111111011111  078IXA
    95 +11111111111111011111  095IXD
   114 +11111111111111111011  114IXB
   120 +11111111111111111110  120IXA
   122 +11111111111111111110  122IXH
   157 +11111111111111111101  157IXA
   174 +11111111111111110111  174IXE
   ...  ...  ...  ...  ...  ...  .   ..  ..
   ...  ...  ...  ...  ...  ...  .   ..  ..
   ...  ...  ...  ...  ...  ...  .   ..  ..
   143 +11111111111001101001  143IXF
   148 +11110110011111101101  148IXG
   151 +11110111111100111100  151IXA

   172 +10011111110101011000  172IXC
   176 +11000110011111110010  176IXF
   207 +11111111001001001001  207IXB
    36 +01111010110001110100  036IXE
    83 +01011101101111000001  083IXA
    88 +11111111001010001001  088IXC
    87 +11111111001010001001  087IXC
   230 +11101111000010001110  230IXC
    52 +11111001101100000001  052IXG
    94 +10111111110100000000  094IXE
   204 +11011011111100000000  204IXF
   159 +01101110001101000000  159IXD
   180 +11111110000000100000  180IXG
    30 +00110000000110000110 1 030IXC
    37 +10001101010000100001  037IX
    41 +01000010110000110001  041IXD
    55 +10001001100101000100  055IXD
    81 +11101001000001000100  081IXH
    89 +11001110000010001000  089IXF
   179 +10110010000100000101  179IXE
   211 +01111000010001100000  211IXE
    18 +11101000000000000101  018IXF
    31 +11001010010000010000  031IXG
   152 +10010000011000000101  152IXG
   173 +00001110000011010000  173IXG
   209 +11001000100000001010  209IXE
   115 +10100000110001000000  115IXE
   123 +10000101001100000000  123IXD
    61 +00110000000001001000  061IXC
   108 +10000000001110000000  108IXA
   190 +00010000010000101000  190IXB
   195 +00000100000000101001 0 195IXA
   196 +10000101000000010000  196IXG
    29 +01000000000000110000  029IXD
   124 +10000000100000100000  124IXD
    32 +10001000000000000000  032IXG
    93 +00001000000000001000  093IXH
```

Figure 4. The Guttman scalogram of the students' answers.

Besides sorting the students' abilities from the highest to the lowest, the Guttman scalogram also sorts them from the easiest to the hardest. The item on the lowest left was the easiest, while that on the top right was the most difficult. Further analysis shows the students who were not careful when working on the examination. In the green highlight table, students with initials 075IXF, 078IXA, and 095IXD only answered one item incorrectly out of 20 given. They worked on the last five items that had higher difficulty indicating that their ability was sufficient, though they were not careful enough to answer the easier items. Meanwhile, the students with initials 056IXF and 065IXF in the blue highlight box answered the first 19 items correctly but incorrectly answered item number 20, showing their ability was still insufficient for the question. The Guttman scalogram can also detect students that cheat when working on the Online Mathematics Final Examination. In the black box highlight case, students with initials 088IXC and 087IXC had consecutive numbers, similar math test scores, and patterns in answering the examination questions. From the data, the teacher can further analyze and determine whether these two students cheated.

The Guttman scalogram data is very beneficial for teachers and schools. Teachers and supervisors should ensure the Online Mathematics Final Examination is conducted honestly and a pure final score without cheating is obtained. They can also remind students that attitude is more important than grades in the world of work.

In addition, the Guttman scalogram data can be used by teachers to classify students' abilities. By applying this data on the Rasch model to analyze the daily or midterm test, the teacher can identify the items that are still difficult and need to be explained again. The teachers can also assess their low mathematical ability students to assist them and to focus more on helping students with other learning methods or approaches. Meanwhile, the teachers can provide more challenging items for the high-ability students to improve their other math skills or prepare them for competitions.



```
    DICHOTOMOUS CURVES
P      -+--------------+--------------+--------------+--------------+-
R  1.0 +                                                             +
O      |                                                             |
B      |0                                                           1|
A      | 000000                                               111111 |
B   .8 +       00000                                     11111       +
I      |            0000                             1111            |
L      |                0000                     1111                |
I      |                    000               111                    |
T   .6 +                       000         111                       +
Y      |                          000   111                          |
    .5 +                             ***                             +
O      |                          111   000                          |
F   .4 +                       111         000                       +
       |                    111               000                    |
R      |                1111                     0000                |
E      |            1111                             0000            |
S   .2 +       11111                                     00000       +
P      | 111111                                               000000 |
O      |1                                                           0|
N      |                                                             |
S   .0 +                                                             +
E      -+--------------+--------------+--------------+--------------+-
       -2             -1              0              1              2
```

Figure 5. Probability of response - dichotomous curves.

Figure 5 illustrates the probability categories and shows that the logit peak was around 0.9, and numbers 1 and 0 did not cover each other. The correctly and incorrectly answered items were balanced, meaning the test adequately measured and divided the students' mathematical ability and into the low and high categories. Consequently, the items can be concluded to have appropriately divided the students' abilities.

4. Discussion

The final examinations in grades 9 and 12 are very important for determining the students' ability level in schools, cities, and provinces. According to the results, the government can plan to improve the schools' quality in each province. The schools can measure their students' abilities quality, the teachers can make evaluations on lesson plans and teaching methods, and the goal of improving the quality of education in Indonesia for Teacher training in sustainable education can be achieved. However, the government has not held a national examination for two years, particularly in 2020 and 2021, and has entrusted each school to make its final examination to determine students' graduation. Therefore, final examination items should be made appropriately and professionally and each item should be carefully evaluated to ensure the students' mathematical abilities are measured properly. In addition, the items should be fairly moderate at difficult level to allow the students to pass with good grades and slightly difficult to test their abilities.

Subsequently, the research result shows the importance of validating the final examination items. It indicates that although public schools had the predicate "A," the construct validity results did not reach the minimum score. Furthermore, the Rasch model found several items that were not good, and 4 out of the 20 items needed to be evaluated and revised because they were outside the fit item criteria. Hence, these 4 items require further revision and analysis to correspond with the fit category.

The contribution of this research is the importance of preparing the final examination items appropriately and seriously. It also explains the steps of using the Rasch model to analyze important information needed for preparing the final examination. First, determine whether the construct validity is in a good category (X>40%). Second, investigating the misfits on the statistical item table to observe the difficulty level of the items as a whole and analyze the personal statistics to evaluate the students' overall abilities. Third, using the Wright Maps to assess a comparison between the abilities of each student and the difficulty level of each item in more detail. Fourth, analyzing the Guttman scalogram to observe the students' ability to answer the items in detail. In conclusion, these steps ensure that all items are well-prepared and checked whether the test is unidimensional. It also ascertains that there is no sentence error evidence, luck in answering, distractor analysis error, miscoding, etc. This research concluded that the Rasch measurement model can be used by students, teachers, and schools to adequately prepare the final examination items.

5. Conclusions and Implications

The Covid-19 pandemic has made the world educational institutions to conduct learning online [31], [32]. This situation made Indonesia stop the national examinations, causing each school to prepare its final examination items for determining the students' graduation. By processing the final examination data using the Rasch measurement model, the items and the students' work can be analyzed. The model can show data and help teachers and schools analyze the students' abilities in working on and answering each test item on the online mathematics final examination. Furthermore, the Rasch model shows the difficulty level of each item as an evaluation to improve the quality of the items properly. Therefore, this model can enhance the achievement of item results and determine better goals, meaning the online final examination distractor analysis in mathematics courses works well.

Also, the findings show that the Rasch model can analyze the quality of the Online Mathematics Final Examination Items at public school in Indonesia. The results of the overall construct validation show that the items were in the accepted category, and after further investigation, four items out of the 20 were discovered not to meet the fitness criteria. This means the four items need to be further observed, revised, or discarded. Also, the results of the person reliability show in the good category and item reliability show that the items were in the 'very good' category. The students' abilities to work on the online final examination was above average. In addition, the *Guttman scalogram of responses* can show data on the student abilities from the highest to the lowest, as well as the less careful students, the ability limits, and those who likely cheated.

The research suggested that the teachers should prepare well to achieve good construct validity when examining students, especially on important examinations, such as UTS and UAS. It is also recommended that teachers use the Rasch model to obtain more specific items analysis. The model can help produce more objective examination items that can work well to measure the students' abilities. Another important message is that teachers should focus on students with low mathematical abilities and help them improve their learning outcomes. With Rasch analysis, the items and sub-topics that students don't understand can be discovered fully, and the teachers can use new methods or ways to re-explain the topics.

6. Limitations and suggestions for further research

There were several limitations in this research. First, it was only conducted with a small scale of 234 students, and only one public school with a good predicate in Bandung city was selected, hence the preparation of the final examination items for the 9th-grade junior high school had good results. However, analysis of public schools in remote areas may get different results, and further research can analyze the final examinations in private schools. The second limitation is that only the final examination items at the junior high school were analyzed, hence other school levels need to be investigated. Third, only mathematics examination items were analyzed, while the final examination to determine student graduation consists of many subjects such as science, foreign language, and others. Therefore, further research can analyze other subjects tested on the final examination using the Rasch model.

Funding: this research funding by Beijing Normal University.

Acknowledgments: Thanks for public school In Indonesia who support this data.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] T. T. Wijaya, Z. Ying, A. Purnama and N. Hermita, “Indonesian students’ learning attitude towards online learning during the coronavirus pandemic,” *Psychol. Eval. Technol. Educ. Res.*, vol. 3, no. 1, pp. 17–25, 2020, doi: 10.33292/petier.v3i1.56.

[2] T. T. Wijaya, Z. Ying and L. Suan, “Gender and Self-regulated Learning During COVID-19 Pandemic in Indonesia,” *J. basicedu*, vol. 4, no. 3, pp. 725–732, 2020, doi: 10.31004/basicedu.v4i3.422.

[3] X. Wu, Y. Zhou and Z. Mo, “A Comparative Study on the comprehensive difficulty of Junior High School National Examination,” *J. Educ.*, vol. 2, no. 4, pp. 352–366, 2020.

[4] S. O. Manullang, E. Satria, U. Krisnadwipayana and U. B. Hatta, “The Review of the International Voices on the Responses of the Worldwide School Closures Policy Searching during Covid-19 Pandemic,” vol. 5, no. 2, pp. 1–13, 2020.

[5] H. H. Hohne, “the Prediction of Academic Success,” *Aust. J. Psychol.*, vol. 1, no. 1, pp. 38–42, 1949, doi: 10.1080/00049534908256014.

[6] A. Purwanto, R. Pramono, M. Asbari, P. B. Santoso, L. M. Wijayanti, C. C. Hyun and P. R. S, “Studi Eksploratif Dampak Pandemi COVID-19 Terhadap Proses Pembelajaran Online di Sekolah Dasar,” *EduPsyCouns J.*, vol. 2, no. 1, 2020.

[7] V. Locastro, “The English in Japanese university entrance examinations: a sociocultural analysis,” *World Englishes*, vol. 9, no. 3, pp. 343–354, 1990, doi: 10.1111/j.1467-971X.1990.tb00271.x.

[8] C. Bai, W. Chi and X. Qian, “China Economic Review Do college entrance examination scores predict undergraduate GPAs ? A tale of two universities ☆,” *China Econ. Rev.*, vol. 30, pp. 632–647, 2014, doi: 10.1016/j.chieco.2013.08.005.

[9] A. G. Sanz and A. S. Pavón, “Toward implementing computer-assisted foreign language assessment in the official Spanish University Entrance Examination,” in *Proceedings of the 2015 EUROCALL Conference, Padova, Italy*, 2015, pp. 215–220.

[10] W. T. Tseng, H. F. Cheng and X. Gao, “Validating a motivational self-guide scale for language learners,” *Sustain.*, vol. 12, no. 16, 2020, doi: 10.3390/su12166468.

[11] J. A. Weller, N. F. Dieckmann, M. Tusler, C. K. Mertz, W. J. Burns and E. Peters, “Development and Testing of an Abbreviated Numeracy Scale: A Rasch Analysis Approach,” *J. Behav. Decis. Mak.*, vol. 26, no. 2, pp. 198–212, 2013, doi: 10.1002/bdm.1751.

[12] J. Ahmad and N. M. Siew, “Curiosity towards stem education: A questionnaire for primary school students,” *J. Balt. Sci. Educ.*, vol. 20, no. 2, pp. 289–304, 2021, doi: 10.33225/jbse/21.20.289.

[13] J. A. Martínez-gonzález, V. T. Díaz-padilla and E. Parra-lópez, “Study of the tourism competitiveness model of the world economic forum using rasch’s mathematical model: The case of portugal,” *Sustain.*, vol. 13, no. 13, 2021, doi: 10.3390/su13137169.

[14] S. E. Mokshein, H. Ishak and H. Ahmad, “The use of rasch measurement model in English testing,” *Cakrawala Pendidik.*, vol. 38, no. 1, pp. 16–32, 2019, doi: 10.21831/cp.v38i1.22750.

[15] J. Niens, L. Richter-Beuschel, T. C. Stubbe and S. Bögeholz, “Procedural knowledge of primary school teachers in madagascar for teaching and learning towards land-use-and health-related sustainable development goals,” *Sustain.*, vol. 13, no. 16, 2021, doi: 10.3390/su13169036.

[16] S. C. Yang, M. Y. Tsou, E. T. Chen, K. H. Chan and K. Y. Chang, "Statistical item analysis of the examination in anesthesiology for medical students using the Rasch model," *J. Chinese Med. Assoc.*, vol. 74, no. 3, pp. 125–129, 2011, doi: 10.1016/j.jcma.2011.01.027.

[17] A. Faradillah and L. Febriani, "Mathematical Trauma Students' Junior High School Based on Grade and Gender," *Infin. J.*, vol. 10, no. 1, p. 53, 2021, doi: 10.22460/infinity.v10i1.p53-68.

[18] H. Othman, I. Asshaari, H. Bahaludin, Z. M. Nopiah and N. A. Ismail, "Application of Rasch Measurement Model in Reliability and Quality Evaluation of Examination Paper for Engineering Mathematics Courses," *Procedia - Soc. Behav. Sci.*, vol. 60, no. 2009, pp. 163–171, 2012, doi: 10.1016/j.sbspro.2012.09.363.

[19] O. Karlin and S. Karlin, "Making Better Tests with the Rasch Measurement Model," *InSight A J. Sch. Teach.*, vol. 13, no. 1, pp. 76–100, 2018, doi: 10.46504/14201805ka.

[20] H. Shamsuddin and A. Z. Khairani, *Proceedings of the Regional Conference on Science, Technology and Social Sciences (RCSTSS 2016)*, no. Rcstss 2016. Springer Singapore, 2019.

[21] L. Huang, F. Huang, P. T. Oon and M. C. K. Mak, "Constructs evaluation of student attitudes towards science," *Eurasia J. Math. Sci. Technol. Educ.*, vol. 15, no. 12, 2019, doi: 10.29333/ejmste/109168.

[22] R. L. Haspel, Y. Lin, P. Fisher, A. Ali and E. Parks, "Development of a validated exam to assess physician transfusion medicine knowledge," *Transfusion*, vol. 54, no. 5, pp. 1225–1230, 2014, doi: 10.1111/trf.12425.

[23] H. Retnawati and N. F. Wulandari, "The development of students' mathematical literacy proficiency," *Probl. Educ. 21st Century*, vol. 77, no. 4, pp. 502–514, 2019, doi: 10.33225/pec/19.77.502.

[24] M. Tavakol and R. Dennick, "Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72," *Med. Teach.*, vol. 35, no. 1, pp. e838–e848, 2013, doi: 10.3109/0142159X.2012.737488.

[25] N. B. Jennings, M. K. Slack, L. E. Mollon and T. L. Warholak, "Measurement characteristics of a concept classification exam using multiple case examples: A Rasch analysis," *Curr. Pharm. Teach. Learn.*, vol. 8, no. 1, pp. 31–38, 2016, doi: 10.1016/j.cptl.2015.09.010.

[26] Z. M. Nopiah, N. A. Ismail, H. Othman, I. Asshaari, N. Razali, M. H. Osman and M. H. Jamalluddin, "Identification of student achievement and academic profile in the linear algebra course: An analysis using the Rasch model," *2011 3rd Int. Congr. Eng. Educ. Rethink. Eng. Educ. W. Forward, ICEED 2011*, pp. 197–202, 2011, doi: 10.1109/ICEED.2011.6235389.

[27] G. Engelhard and J. Wang, "Developing a concept map for rasch measurement theory," *Springer Proc. Math. Stat.*, vol. 322, pp. 19–29, 2020, doi: 10.1007/978-3-030-43469-4_2.

[28] B. Sumintono and W. Widhiarso, "Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial," 2013.

[29] W. J. Boone, M. S. Yale and J. R. Staver, *Rasch analysis in the human sciences*. Springer, 2014.

[30] J. Linacre, *A user's guide to WINSTEPS: Rasch model computer programs*. MESA Press, Chicago, 2012.

[31] X. Zhu and J. Liu, "Education in and After Covid-19: Immediate Responses and Long-Term Visions," *Postdigital Sci. Educ.*, 2020, doi: 10.1007/s42438-020-00126-3.

[32] M. S. Alabdulaziz, "COVID-19 and the use of digital technology in mathematics education," *Educ. Inf. Technol.*, no. 0123456789, 2021, doi: 10.1007/s10639-021-10602-3.