JRERE 13 (2) 2024 169-179

**Journal of Research and**
**Educational Research Evaluation**

http://journal.unnes.ac.id/sju/index.php/jere

# Evaluating Validity and Reliability of Scientific Literacy Among Indonesian Secondary Students in PISA 2015 Context: Rasch Model Analysis

**Winata Tegar Saputra[1]⊠, Nuryani Y. Rustaman[2], Lilit Rusyati[3]**

[1,2,3]Department of Science Education, Faculty Mathematics and Science Education,
Universitas Pendidikan Indonesia, Bandung, Indonesia

**Article Info**

History Articles
Received:
12 May 2024
Accepted:
10 June 2024
Published:
30 August 2024

Keywords:
*Scientific literacy, rasch model, validity, reliability instrument*

## Abstract

Indonesia has followed the Scientific Literacy Test in PISA OECD and the rank is at the bottom. Scientific literacy is important things in science education because it involves the skills of scientific thinking in society life. This study aims to find the validity and reliability of scientific literacy in PISA 2015 using the Rasch model. The method of this research used quantitative survey research that assessed 15 years of 92 secondary school students. The validity in general shows the test items are not valid, but the validity in each test item shows 17 of 20 questions are valid. Reliability in person is lower than reliability in items. Even though, the test items still have good quality from the reliability findings. The novelty of this study is the validity and reliability of the PISA 2015 test items in Indonesian students as the sample. PISA 2015 used scientific literacy as the dominant test at that time. This study has benefitted other researchers by showing the importance of the validity and reliability of the instruments and ensuring the quality of the instruments for other researchers who will conduct future research using good validity and reliable instruments.

⊠Correspondence Address :
Department of Science Education, Faculty Mathematics and Science
Education, Universitas Pendidikan Indonesia, Bandung, Indonesia,
40154
E-mail : winatategar08@upi.edu

**p-ISSN 2252-6420**
**e-ISSN 2503-1732**

## INTRODUCTION

Scientific literacy become the trend in science education. The application of the skills or competencies that students acquire in school and later throughout their lives as contributing members of society is how literacy is made visible (Queiruga-Dios et al., 2020). According Norris and Phillips (2003) one of the terms in scientific literacy is the ability to think critically about science and to deal with scientific expertise and the ability of the student to think scientifically. Scientific literacy is posing questions about nature and using science to uncover the answers (Sari et al., 2024). Scientific literacy also developed into Civic Scientific Literacy which assesses scientific knowledge, scientific method, problem-solving, and scientific thought and spirit of science (Liu et al., 2024)

Many countries assess the ability of students to assess their scientific literacy by the OECD Programme for International Student Assessment (PISA). OECD (2015) state there are three categories of scientific knowledge that are necessary for pupils to be able to exhibit these competencies. These are described as content knowledge, procedural knowledge, and epistemic knowledge. The PISA program evaluates 15-year-olds in three areas: scientific, mathematical, and reading literacy. It is administered every three years (Harlen, 2001).

Indonesia followed the PISA. The PISA 2018 literacy results for Indonesia showed a significant decrease from the PISA 2015 results (Amini & Sinaga, 2021) and the rank is improving in 2022 than 2018. However, Indonesian rank in literacy is still at the bottom. PISA test units are not aligned with any curriculum utilized in schools in several parts of Indonesia Wati et al (2017). Also, the interest of the Indonesian researcher is still low (Ni'mah, 2019). The goal of teaching is for students to gain value based on the minimal mastering requirements, removing any doubt about the competency of science students (Basam et al., 2017).

The PISA 2015 was developed and analyzed by Rasch Model Measurement to find out the quality and difficulty of the test item (Sihombing et al., 2019). The quality of the test item was analyzed by using Item Response Theory (IRT) and Classic Test Theory (CTT) (Arabbani et al., 2019). According Darman and Colleagues (2024) prospective teacher students' scientific inquiry literacy can be precisely assessed using the Scientific Inquiry Literacy Instrument (SILI). The quality of the development scientific literacy test instrument used the Rasch Model Approach (Nurul et al., 2023). The development of PISA scientific literacy used to monitor the change in education stages (Zhang et al., 2023). Using the 2015 PISA features and indicators as a guide, test instruments for basic material scientific literacy were created (Wati et al., 2023). The Rasch model analyzes item fit statistics, dimensionality, person-item mappings (Wright maps), differential item functioning (DIF), and item category structure in developing the instrument of energy literacy (Yusup, 2021). According Hastuti and Colleagues (2022) a reliable tool for assessing scientific literacy within the context of inquiry-based learning that incorporates ethnoscience.

Analyze the validity, reliability, discriminatory index, and item difficulty level of scientific literacy (Yusmaita et al., 2022). The Rasch model measured learning objectives for pupils to ascertain their actual ability (Darmana et al., 2021). The instrument is adapted from PISA 2015 test items using the Rasch model. The reason that PISA 2015 was used as the instrument was that PISA 2015 has majored in scientific literacy (OECD, 2017). A key element of high-quality research is the use of valid and reliable tests or instruments to measure these kinds of constructs (Kimberlin & Winterstein, 2008). According to Ahmed and Ishtiaq (2021), The constancy of a method when measuring something is called reliability. The degree to which a methodology measures a variable that it is intended to measure is what is known as validity. Therefore, this study aims to identify whether the PISA 2015 test items are valid and reliable to Indonesian Students at this moment.

## METHODS

The method in this study used quantitative survey research (Cresswell, 2012). This study assesses 15 years for 92 secondary school students, that appropriate with the OECD requirement. Test items used PISA 2015 as the instrument and analysis by Rasch Model To present diverse types of evidence related to construct validity (Lim et al., 2009). According Bond (2003) The validity of the testing process is then directly demonstrated by the Rasch measurement indicators of item order and item fit, particularly when the test material is expressly included in substantive theory about the construct being studied. The value of validity and reliability is important for other researchers as showing the quality of the instrument (Ghazali, 2016).

Table 1. shows the analysis of findings in the Rasch Model state by Sumintono and Widhiarso (2015) and Darmana and Colleagues (2021). This analyzes all of the research questions in this study. Therefore, we have the implication for further research to improve scientific literacy in Indonesia that is suitable for the cognitive demand of OECD PISA in the future.

After analyzing the Rasch model in validity and reliability. The result could categorize the validity in general and item, as well as reliability in person and item into the interpretation Table 1. The result of the analysis of the validity and reliability of the Rasch model is explained in the result and discussion.

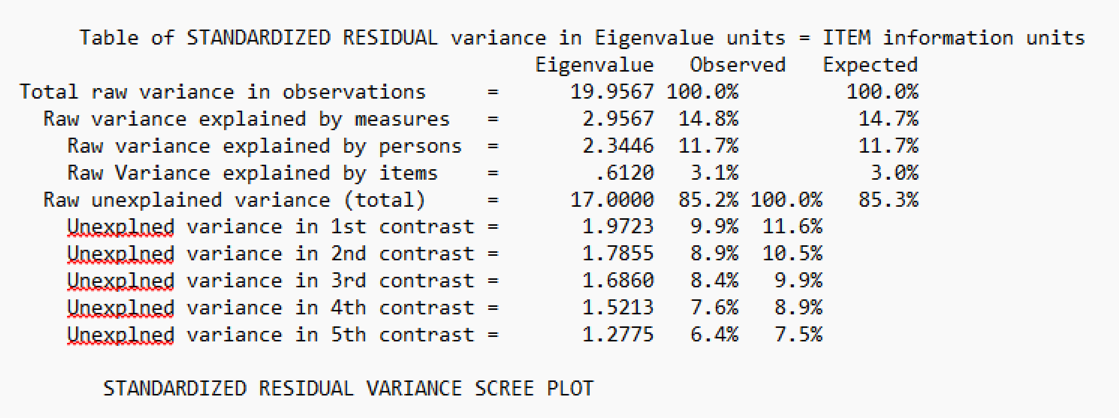
## RESULTS AND DISCUSSION

### Validity of the PISA 2015 Test Items

According Razali and Colleagues (2016) the level to which a study testing instrument acquires the intended data. There are a lot of types of validity. First, the validity of content needs to the judged by the expert, and the validity of question items that analyzed by the Rasch model. While the validity of the construct is shown Rasch model is explained in Figure 1.

**Table 1.** Criteria and aspects in different view

| Statistics | Fit Indices | Interpretation |
|---|---|---|
| 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 High |
| 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 | | A high separation value suggests that the instrument is of high quality because it can distinguish between the respondent and the item group. |
| Item difficulty | >1 | Very difficult |
| | 0.5 ≤ b < 1 | Difficult |
| | - 0.5 ≤ b < 0.5 | Moderate |
| | - 0.5 ≤ b < -1 | Easy |
| | B ≤ -1 | Very easy |



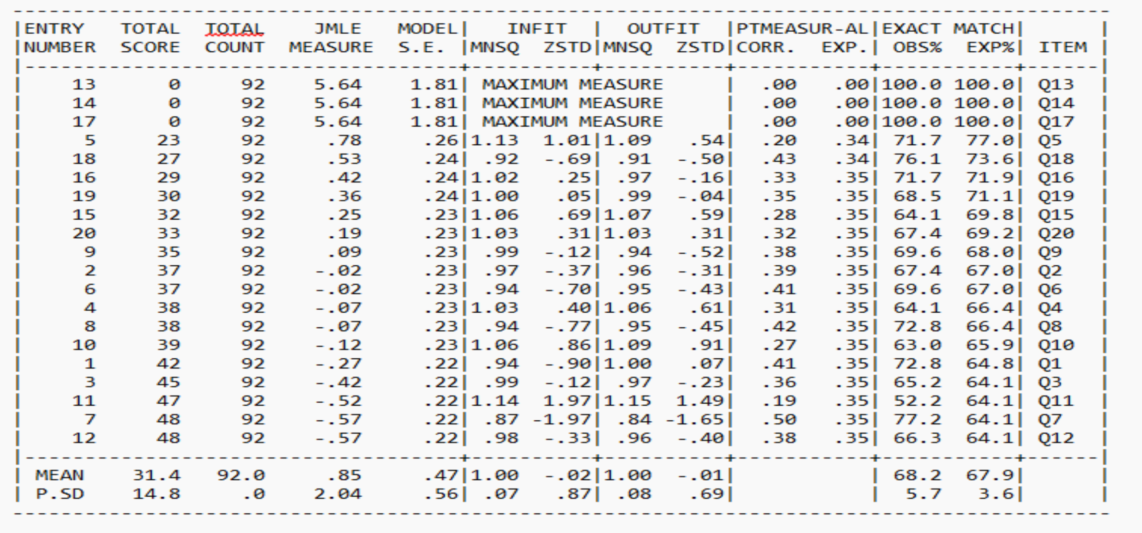
```
    Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units
                                              Eigenvalue   Observed     Expected
Total raw variance in observations      =      19.9567 100.0%         100.0%
  Raw variance explained by measures    =       2.9567  14.8%          14.7%
    Raw variance explained by persons   =       2.3446  11.7%          11.7%
    Raw Variance explained by items     =        .6120   3.1%           3.0%
  Raw unexplained variance (total)      =      17.0000  85.2% 100.0%   85.3%
    Unexplned variance in 1st contrast =       1.9723   9.9%  11.6%
    Unexplned variance in 2nd contrast =       1.7855   8.9%  10.5%
    Unexplned variance in 3rd contrast =       1.6860   8.4%   9.9%
    Unexplned variance in 4th contrast =       1.5213   7.6%   8.9%
    Unexplned variance in 5th contrast =       1.2775   6.4%   7.5%

      STANDARDIZED RESIDUAL VARIANCE SCREE PLOT
```

**Figure 1.** The result of the Rasch validity in general

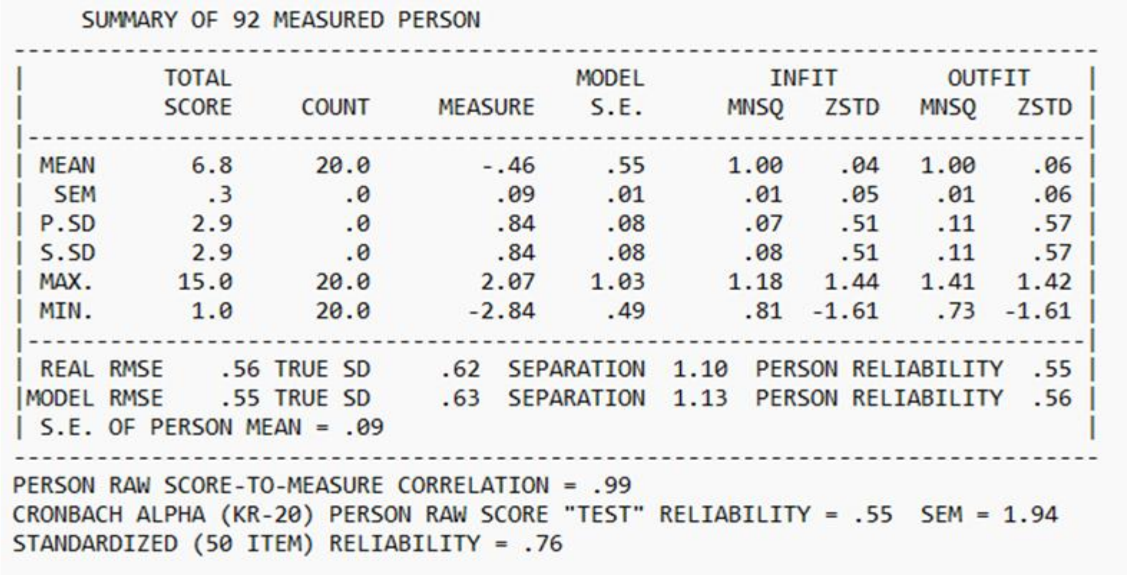
Figure 1 explains the validity of general test items. We could see in the result of Raw Variance Explained by Measure is 14.8%. It is under 20% which means the validity is poor. However, if one looks at each item. Some items have validity. We can see this in Figure 2. As we analyze in Outfit MNSQ, the validity has a value from 0.5 up to 1.5. it means the test items that are valid are all of the questions except question 13, question 14, and question 17. It is because the test items in invalid test items are complex multiple choice. It means the answers could be more than 1. In addition, if the students only give one answer, it will be incorrect even if they choose one of the correct answers.

As we know, Indonesia is still at the bottom of the rank in the PISA test. The reason is that Indonesian students are not familiar with the PISA test. According Nugrahanto and Zuchdi (2019) there are several factors that influence scientific literacy achievement: (1) the roles of the school, (2) the differences between public and private schools, and (3) the background in socio-economic. Also, Fenanlampir et al (2019) state there are several factors that influence the achievement by Indonesian Students is still low. They state (1) limited learning facilities, (2) access to the educational site, (3) equity education for educators, and (4) Indonesian islands-based. Argina et al (2017) state the factors of the stagnant achievement in PISA are: (1) education funds, (2) equity and quality of Teachers, (3) education system; and (4) decentralization of education.



```
----------------------------------------------------------------------------------------------
|ENTRY   TOTAL  TOTAL    JMLE   MODEL|   INFIT    |  OUTFIT   |PTMEASUR-AL|EXACT  MATCH|      |
|NUMBER  SCORE  COUNT  MEASURE  S.E. |MNSQ   ZSTD|MNSQ   ZSTD|CORR.  EXP.| OBS%  EXP%| ITEM |
|------------------------------------+-----------+-----------+-----------+-----------+------|
|    13      0     92    5.64   1.81| MAXIMUM MEASURE        |  .00   .00|100.0 100.0| Q13  |
|    14      0     92    5.64   1.81| MAXIMUM MEASURE        |  .00   .00|100.0 100.0| Q14  |
|    17      0     92    5.64   1.81| MAXIMUM MEASURE        |  .00   .00|100.0 100.0| Q17  |
|     5     23     92     .78    .26|1.13   1.01|1.09    .54|  .20   .34| 71.7  77.0| Q5   |
|    18     27     92     .53    .24| .92   -.69| .91   -.50|  .43   .34| 76.1  73.6| Q18  |
|    16     29     92     .42    .24|1.02    .25| .97   -.16|  .33   .35| 71.7  71.9| Q16  |
|    19     30     92     .36    .24|1.00    .05| .99   -.04|  .35   .35| 68.5  71.1| Q19  |
|    15     32     92     .25    .23|1.06    .69|1.07    .59|  .28   .35| 64.1  69.8| Q15  |
|    20     33     92     .19    .23|1.03    .31|1.03    .31|  .32   .35| 67.4  69.2| Q20  |
|     9     35     92     .09    .23| .99   -.12| .94   -.52|  .38   .35| 69.6  68.0| Q9   |
|     2     37     92    -.02    .23| .97   -.37| .96   -.31|  .39   .35| 67.4  67.0| Q2   |
|     6     37     92    -.02    .23| .94   -.70| .95   -.43|  .41   .35| 69.6  67.0| Q6   |
|     4     38     92    -.07    .23|1.03    .40|1.06    .61|  .31   .35| 64.1  66.4| Q4   |
|     8     38     92    -.07    .23| .94   -.77| .95   -.45|  .42   .35| 72.8  66.4| Q8   |
|    10     39     92    -.12    .23|1.06    .86|1.09    .91|  .27   .35| 63.0  65.9| Q10  |
|     1     42     92    -.27    .22| .94   -.90|1.00    .07|  .41   .35| 72.8  64.8| Q1   |
|     3     45     92    -.42    .22| .99   -.12| .97   -.23|  .36   .35| 65.2  64.1| Q3   |
|    11     47     92    -.52    .22|1.14   1.97|1.15   1.49|  .19   .35| 52.2  64.1| Q11  |
|     7     48     92    -.57    .22| .87  -1.97| .84  -1.65|  .50   .35| 77.2  64.1| Q7   |
|    12     48     92    -.57    .22| .98   -.33| .96   -.40|  .38   .35| 66.3  64.1| Q12  |
|------------------------------------+-----------+-----------+-----------+-----------+------|
| MEAN    31.4   92.0     .85    .47|1.00   -.02|1.00   -.01|           | 68.2  67.9|      |
| P.SD    14.8     .0    2.04    .56| .07    .87| .08    .69|           |  5.7   3.6|      |
----------------------------------------------------------------------------------------------
```

**Figure 2.** The result of Rasch's validity in each test item



```
     SUMMARY OF 92 MEASURED PERSON
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE     COUNT     MEASURE    S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN       6.8      20.0        -.46     .55      1.00    .04   1.00    .06 |
|  SEM        .3        .0         .09     .01       .01    .05    .01    .06 |
| P.SD       2.9        .0         .84     .08       .07    .51    .11    .57 |
| S.SD       2.9        .0         .84     .08       .08    .51    .11    .57 |
| MAX.      15.0      20.0        2.07    1.03      1.18   1.44   1.41   1.42 |
| MIN.       1.0      20.0       -2.84     .49       .81  -1.61    .73  -1.61 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .56 TRUE SD     .62  SEPARATION  1.10  PERSON RELIABILITY  .55 |
|MODEL RMSE    .55 TRUE SD     .63  SEPARATION  1.13  PERSON RELIABILITY  .56 |
| S.E. OF PERSON MEAN = .09                                                   |
-------------------------------------------------------------------------------
PERSON RAW SCORE-TO-MEASURE CORRELATION = .99
CRONBACH ALPHA (KR-20) PERSON RAW SCORE "TEST" RELIABILITY = .55  SEM = 1.94
STANDARDIZED (50 ITEM) RELIABILITY = .76
```

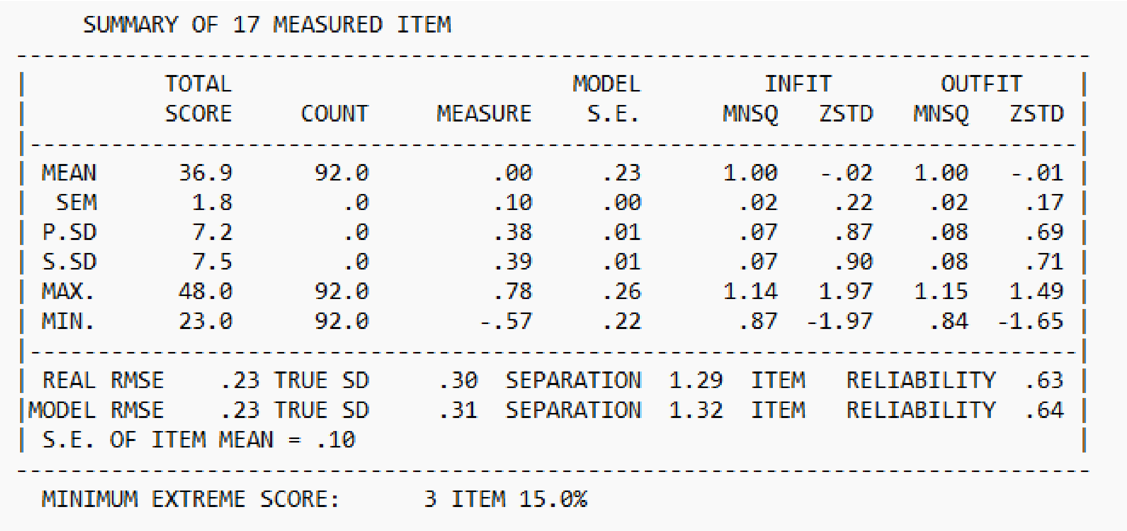
**Figure 3.** The result of Rasch reliability in person

Findings validity has many ways. There are differences between construct validity and content validity. The content validity needs to the judged by the expert for the instrument using Aiken that state if it is less than 0.6 the validity is poor (Setiawan et al., 2020). The content validity used inter-rater agreement that has 81% in developing the instrument (Relkin et al., 2020). The construct validity assesses the instrument by comparing the calculation correlation with the r table (Rery & Erna, 2020). Finding both content validity and construct validity by comprehension validity and the Kaiser-Meyer-Olkin test also could be conducted (Gómez-García et al., 2020). Rodriguez-Macaya and Colleagues (2021) conducted the validity test for the social skills of students in The Kaiser-Meyer-Olkin Test with 0.906 and Bartlett's test. According to Abbas and Sagsan (2020), Convergent validity (the degree to which the scale correlates with other measures measuring comparable dimensions) and discriminant validity (the degree to which they do not correlate with distinct measures) were used to examine the construct validity.

Besides content and construct validity, Barak and Colleagues (2020) analyze another validity: (1) known-groups validity by analysis of the Variations among subgroups, and (2) concurrent validity that is analyzed by triangulation with real practice. The content validity was also analyzed by comparing the scale level content validity index and Cronbach's Alpha, which has a high score two of them (Suneja, 2023). For the next step of validation, convergent validity used the Confirmatory Factor Analysis (CFA) of the sus. inability consciousness questionnaire (SCQ-S) (Olsson et al., 2020). Also, Analyses of play interactions captured on film will be used further to investigate the validity of the teachers' assessments (Sedem et al., 2022). The Validity of the Computational Thinking Scale was analyzed further to clarify the dimension of the validity by using the principal axis analysis and inter-correlation in each dimension (Tsai et al., 2021).

### Reliability of the PISA 2015 Test Items

In assessing 15 years for 92 secondary school students. The test items are shown in the Appendix. The correct answer gives a score of 1 if it is correct and 0 if it is incorrect because this study chose the simple multiple choice and complex multiple choice test items to simplify the analysis by the Rasch model. Figure 3. shows the summary of 92 students measured person and Figure 3. shows the summary of the 20 measured items. The purpose is to find the results of the validity and reliability of the Rasch model. Higher item reliability shows that an item is more successful in differentiating between people with various skills, whereas higher person dependability suggests that a person's replies are more consistent (Wright & Masters, 1982). It shows the Person reliability is 0.55 and 0.56. Based on Table 1. the reliability interpretation is in the Low-Reliability Category in person. Low reliability denotes the model's incapacity to precisely capture the fundamental characteristics or aptitude of the subjects being assessed. According to Downing (2004) low reliability means that significant differences in scores can be anticipated when retesting, and inconsistent assessment results make it harder or impossible to understand the results, which weakens the validity of evidence properly. Also, show inadequate items or a limited range of personal measures (i.e., not enough people with more extreme talents, both high and poor) (Cordier et al., 2018). Low reliability in person has a big impact on the validity of the test results. It could impact inaccurate judgment about individual abilities and not sufficiently accurate to differentiate between high and low achievers. However, winstep resources state that test items may be added.



```
     SUMMARY OF 17 MEASURED ITEM
-------------------------------------------------------------------------------
|          TOTAL                         MODEL         INFIT        OUTFIT    |
|          SCORE      COUNT     MEASURE   S.E.      MNSQ   ZSTD   MNSQ   ZSTD |
|-----------------------------------------------------------------------------|
| MEAN      36.9       92.0         .00    .23      1.00   -.02   1.00   -.01 |
|  SEM       1.8         .0         .10    .00       .02    .22    .02    .17 |
| P.SD       7.2         .0         .38    .01       .07    .87    .08    .69 |
| S.SD       7.5         .0         .39    .01       .07    .90    .08    .71 |
| MAX.      48.0       92.0         .78    .26      1.14   1.97   1.15   1.49 |
| MIN.      23.0       92.0        -.57    .22       .87  -1.97    .84  -1.65 |
|-----------------------------------------------------------------------------|
| REAL RMSE    .23 TRUE SD     .30  SEPARATION  1.29  ITEM   RELIABILITY  .63 |
|MODEL RMSE    .23 TRUE SD     .31  SEPARATION  1.32  ITEM   RELIABILITY  .64 |
| S.E. OF ITEM MEAN = .10                                                     |
-------------------------------------------------------------------------------
  MINIMUM EXTREME SCORE:      3 ITEM 15.0%
```

**Figure 4.** The result of Rasch reliability in item

Figure 4. shows the reliability of item results after analysis by the Rasch model. It shows the Item Reliability is 0.63 and 0.64. It means the reliability in items shows low reliability same as the Person's Reliability. However, the value of the reliability items is higher than reliability in person. Adi and Colleagues (2022) state that person reliability is scored lower than item reliability, which suggests that although student responses are inconsistent, the instrument's items are of very high quality. Also, if we look at the Cronbach alpha value. The reliability level is 0.55 show the level of reliability is poor (Razali et al., 2016). In general, a smaller reliability coefficient is produced by fewer test items (Schumacker & Smith, 2007).

This finding is in line with Rismawati (2023) that find the reliability of the instrument is low using Aiken's test. Another finding, the person reliability index has a lower level at 0.36 and is different from item reliability that has 0.97 (Boone & Noltemeyer, 2017). It has different findings with Fitriyanto et al., (2019) the instruments created for the suitability of electric power steering media have a 99% validity rate, and the category's instrument stability (reliability) level was sufficient and extremely stable. The reliability is also indirect with the value of the coefficient in Cronbach's Alpha (Amirrudin et al., 2020).

Others find the reliability of the Theory of Planned Behaviour has Cronbach's Alpha in 0.73 and 0.85 and the validity of the items have 0.7-0.8 (Dewi et al., 2022). In developing an instrument of the quality of the theories that found the reliability values are above 0.7 (sufficient) and above 0.8 (good) (Yilmaz, 2022). The low person reliability and level index values were caused by the minimal number of items on the instrument, while the insufficient person sample size caused the poor item reliability and strata index values to establish item difficulty (Shafie et al., 2021). The reliability could be used in qualitative interview data using intercoder reliability and interrater reliability (Cheung & Tai, 2023).

After analyzing the validity and reliability, we could conclude that 17 test items could be used for future research. Table 2 shows the valid test item, which elaborates on the competency and context of scientific literacy in PISA. Table 2 shows the validity of each test item.

**Table 2.** The validity of each test item

| Test item | context | Competency | Validity |
|---|---|---|---|
| Question 1 | Local/national - Environmental Quality | Evaluate and Design Scientific Enquiry | Valid |
| Question 2 | Local/national - Environmental Quality | Explain Phenomena Scientifically | Valid |
| Question 3 | Local/national - Environmental Quality | Interpret Data and Evidence Scientifically | Valid |
| Question 4 | Global - Natural Resources | Explain Phenomena Scientifically | Valid |
| Question 5 | Global - Hazards | Interpret Data and Evidence Scientifically | Valid |
| Question 6 | Global - Hazards | Interpret Data and Evidence Scientifically | Valid |
| Question 7 | Local/national - Hazards | Interpret Data and Evidence Scientifically | Valid |
| Question 8 | Local/national - Hazards | Explain Phenomena Scientifically | Valid |
| Question 9 | Local/ national - Hazards | Explain Phenomena Scientifically | Valid |
| Question 10 | Local/National - Frontiers | Interpret Data and Evidence Scientifically | Valid |
| Question 11 | Local/national - Frontiers | Interpret Data and Evidence Scientifically | Valid |
| Question 12 | Personal – Health and Disease | Explain Phenomena Scientifically | Valid |
| Question 13 | Global - Frontiers | Explain Phenomena Scientifically | Invalid |
| Question 14 | Global - Frontiers | Explain Phenomena Scientifically | Invalid |
| Question 15 | Global - Frontiers | Interpret Data and Evidence Scientifically | Valid |
| Question 16 | Global – Environmental Quality | Explain Phenomena Scientifically | Valid |
| Question 17 | Global – Environmental Quality | Interpret Data and Evidence Scientifically | Invalid |
| Question 18 | Local/national – Environmental Quality | Explain Phenomena Scientifically | Valid |

| Test item | context | Competency | Validity |
|---|---|---|---|
| Question 19 | Local/national – Environmental Quality | Interpret Data and Evidence Scientifically | Valid |
| Question 20 | Personal – Health and Disease | Explain Phenomena Scientifically | Valid |

As we analyze in Table 2, three science competencies are in this test item. There are explained phenomena scientifically, interpret data and evidence, and evaluate and design scientific inquiry. While the context is coming from local/personal into the global. Also, the test items have several topics as we see in Table.

## CONCLUSION

After investigating the research questions, there is a correlation between the validity and reliability of the scientific literacy test items from PISA 2015 when assessing Indonesian students. As the Rasch model analysis result. We could conclude that the validity of the test items in general is not valid. However, if we analyze each item a lot of test items get validity and few of the test items are invalid because all of the students have incorrect answers in the complex multiple choice. Then, in reliability findings also have different categories between the person and items. Item reliability has a higher value than Person reliability.

Determining the validity and reliability of instruments in research implies ensuring that the results obtained are both accurate and consistent. A valid and reliable instrument accurately measures the intended concept or construct and yields consistent results when used multiple times with the same group of participants. It has benefits for further research in assessing the pilot test for good validity and reliability of the instruments. It makes the instrument high quality and has a solid foundation for another researcher to use the instrument for future research.

## REFERENCES

Abbas, J., & Sagsan, M. (2020). Identification of Key Employability Attributes and Evaluation of University Graduates' Performance: Instrument Development and Validation. *Higher Education, Skills and Work-Based Learning*, *10*(3), 449–466.

Adi, N. R. M., Amaruddin, H., Maulana, H., Adi, M., & Laili Qurroti A'yun, I. (2022). Validity and Reliability Analysis Using the Rasch Model to Measure the Quality of Mathematics Test Items of Vocational High Schools. *Journal of Research and Educational Research Evaluation*, *11*(1), 103–113.

Ahmed, I., & Ishtiaq, S. (2021). Reliability and validity: Importance in Medical Research. *Journal of the Pakistan Medical Association*, *71*(10), 2401–2406.

Amini, S., & Sinaga, P. (2021). Inventory of Scientific Literacy Ability of Junior High School Students Based on The Evaluation of PISA Framework Competency Criteria. *Journal of Physics: Conference Series*, *1806*(1).

Amirrudin, M., Nasution, K., & Supahar, S. (2020). Effect of Variability on Cronbach Alpha Reliability in Research Practice. *Jurnal Matematika, Statistika Dan Komputasi*, *17*(2), 223–230.

Arabbani, F. K., Mulyani, S., Mahardiani, L., & Ariani, S. R. D. (2019). Analysis The Quality of Instrument For Measuring Chemical Literacy Abilities of High School Student Using Rasch Model. *AIP Conference Proceedings*, *2194*.

Argina, A. W., Mitra, D., Ijabah, N., & Setiawan, R. (2017). Indonesian PISA Result: What Factors and What Should Be Fixed? *Proceedings Education and Language International Conference*, *1*(1), 69–79.

Barak, M., Watted, A., & Haick, H. (2020). Establishing The Validity and Reliability of A Modified Tool For Assessing Innovative Thinking of Engineering Students. *Assessment and Evaluation in Higher Education*, *45*(2), 212–223.

Basam, F., Rusilowati, A., & Ridlo, S. (2017). Analysis of Science Literacy Learning with Scientific Inquiry Approach in Increasing Science Competence of Students. *Journal of Primary Education*, *6*(3), 174–184.

Bond, T. G. (2003). Validity and assessment : a Rasch measurement perspective. *Metodologia de Las Ciencias Del Comportamiento*, *5*(2), 179–194.

Boone, W. J., & Noltemeyer, A. (2017). Rasch Analysis: A Primer for School Psychology Researchers and Practitioners. *Cogent Education*, *4*(1).

Cheung, K. K. C., & Tai, K. W. H. (2023). The Use of Intercoder Reliability in Qualitative Interview Data Analysis in Science Education. *Research in Science and Technological Education*, *41*(3), 1155–1175.

Cordier, R., Speyer, R., Schindler, A., Michou, E., Heijnen, B. J., Baijens, L., Karaduman, A., Swan, K., Clavé, P., & Joosten, A. V. (2018). Using Rasch Analysis to Evaluate the Reliability and Validity of the Swallowing Quality of Life Questionnaire: An Item Response Theory Approach. *Dysphagia*, *33*(4), 441–456.

Cresswell, J. W. (2012). *Planning, conducting, and evaluating quantitative and qualitative research*.

Darman, D. R., Suhandi, A., Kaniawati, I., Samsudin, A., & Wibowo, F. C. (2024). Development and Validation of Scientific Inquiry Literacy Instrument (SILI) Using Rasch Measurement Model. *Education Sciences*, *14*(3).

Darmana, A., Sutiani, A., Nasution, H. A., Ismanisa*, I., & Nurhaswinda, N. (2021). Analysis of Rasch Model for the Validation of Chemistry National Exam Instruments. *Jurnal Pendidikan Sains Indonesia*, *9*(3), 329–345.

Dewi, N. U., Khomsan, A., Dwiriani, C. M., Riyadi, H., Ekayanti, I., & Nurulfuadi. (2022). Validity and Reliability of The Theory of Planned Behavior Questionnaire on The Balanced Dietary Behavior of Adolescents In A Post-Disaster Area. *Journal of Health Sciences*, *12*(1), 62–73.

Downing, S. M. (2004). Reliability: On The Reproducibility of Assessment Data. *Medical Education*, *38*(9), 1006–1012.

Fenanlampir, A., Batlolona, J. R., & Imelda, I. (2019). The Struggle of Indonesian Students in The Context of TIMSS and Pisa Has Not Ended. *International Journal of Civil Engineering and Technology*, *10*(2), 393–406.

Fitriyanto, J. N., Widjanarko, D., & Khumaedi, M. (2019). Validity and Reliability Test of Assessment Instrument of the Suitability of Electric Power Steering Media. *Journal of Vocational Career Education*, *4*(1), 61–69.

Gómez-García, M., Matosas-López, L., & Ruiz-Palmero, J. (2020). Social Networks Use Patterns among University Youth: The Validity and Reliability of an Updated Measurement Instrument. *Sustainability (Switzerland)*, *12*(9).

Harlen, W. (2001). The Assessment of Scientific Literacy in The OECD/PISA Project. *Studies in Science Education*, *36*(1), 79–104.

Hastuti, P. W., Anjarsari, P., & Yamtinah, S. (2022). Assessment Instrument Scientific Literacy on Addictive Substances Topic in Inquiry Based Learning Integrated Ethnoscience. *Journal of Science Education Research*, *6*(1), 31–36.

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and Reliability of Measurement Instruments Used in Research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284.

Lim, S. M., Rodger, S., & Brown, T. (2009). *Using Rasch Analysis to Establish The Construct Validity of Rehabilitation Assessment Tools*. *16*(5).

Liu, Y., Wang, J., Zhang, Z., Wang, J., Luo, T., Lin, S., Li, J., & Xu, S. (2024). Development and Validation of an Instrument For Measuring Civic Scientific Literacy. *Disciplinary and Interdisciplinary Science Education Research*, *6*(1).

Ni'mah, F. (2019). Research trends of scientific literacy in Indonesia: Where are we? *Jurnal Inovasi Pendidikan IPA*, *5*(1), 23–30.

Norris, S. P., & Phillips, L. M. (2003). How Literacy in Its Fundamental Sense Is Central to Scientific Literacy. *Science Education*, *87*(2), 224–240.

Nugrahanto, S., & Zuchdi, D. (2019). *Indonesia PISA Result and Impact on The Reading Learning Program in Indonesia*. *297*(Icille 2018), 373–377.

Nurul, M., Azizah, L., Ngabekti, S., Saptono, S., & Susilaningsih, E. (2023). Analysis of Junior High School Students ' Scientific Literacy Using the Rasch Model. *International*

*Conference on Science, Education and Technology*, 539–543.

OECD. (2015). PISA 2015 Cognitive Items. *OECD Programme for International Student Assessment 2015*, 1–89. https://www.oecd.org/pisa/test/PISA2015-Released-FT-Cognitive-Items.pdf

OECD. (2017). PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving (Revised Edition). In *OECD Publishing*.

Olsson, D., Gericke, N., Sass, W., & Boeve-de Pauw, J. (2020). Self-Perceived Action Competence For Sustainability: The Theoretical Grounding and Empirical Validation of A Novel Research Instrument. *Environmental Education Research*, *26*(5), 742–760.

Queiruga-Dios, M. Á., López-Iñesta, E., Diez-Ojeda, M., Sáiz-Manzanares, M. C., & Dorrío, J. B. V. (2020). Citizen Science for Scientific Literacy and The Attainment of Sustainable Development Goals in Formal Education. *Sustainability (Switzerland)*, *12*(10).

R U Rery, Y., & Erna, M. (2020). Validity and Reliability of Assessment Instruments for Analytical Thinking Ability and Chemical Literacy in the Colligative Properties. *Journal of Physics: Conference Series*, *1655*(1).

Razali, S. N., Shahbodin, F., Ahmad, M. H., & Mohd Nor, H. A. (2016). Measuring Validity and Reliability of Perception of Online Collaborative Learning Questionnaire Using Rasch Model. *International Journal on Advanced Science, Engineering and Information Technology*, *6*(6), 966–974.

Relkin, E., de Ruiter, L., & Bers, M. U. (2020). TechCheck: Development and Validation of an Unplugged Assessment of Computational Thinking in Early Childhood Education. *Journal of Science Education and Technology*, *29*(4), 482–498.

Rismawati, K. (2023). Analysis of the Validity and Reliability of Indonesian Vocabulary Mastery Instrument Items Using AIKEN'S Model Calculations. *Journal2.Upgris.Ac.Id*, *1*(1), 17–22.

Rodriguez-Macaya, E., Vidal-Espinoza, R., Gomez-Campos, R., & Cossio-Bolaños, M. (2021). Social Skills of Students from Educational Sciences: Validity, Reliability, and Percentiles for Evaluation. *International Journal of Higher Education*, *10*(3), 259.

Sari, I. N., Mahanal, S., & Setiawan, D. (2024). Implementation of A Problem-Based Learning Model Assisted With Scaffolding to Improve Scientific Literacy and Student Cognitive Learning Outcomes. *BIO-INOVED : Jurnal Biologi-Inovasi Pendidikan*, *6*(1), 35.

Schumacker, R. E., & Smith, E. V. (2007). A Rasch Perspective. *Educational and Psychological Measurement*, *67*(3), 394–409.

Sedem, M., Siljehag, E., Allodi, M. W., & Odom, S. L. (2022). Reliability and Validity of a Teacher Impressions Scale to Assess Social Play of Swedish Children in Inclusive Preschools. *Assessment for Effective Intervention*, *48*(1), 52–61.

Setiawan, A., Pusporini, W., & Dardjito, H. (2020). Observation Instrument for Student Social Attitude in Primary Schools: Validity and Reliability. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *24*(1), 76–87.

Shafie, S., Abd Majid, F., Hoon, T. S., & Damio, S. M. (2021). Evaluating Construct Validity and Reliability of Intention to Transfer Training Conduct Instrument Using Rasch Model Analysis. *Pertanika Journal of Social Sciences and Humanities*, *29*(2), 1055–1070.

Sihombing, R. U., Naga, D. S., & Rahayu, W. (2019). A Rasch Model Measurement Analysis on Science Literacy Test of. *Indonesian Journal Educational Review*, *6*(1), 44–55.

Sumintono, B., & Widhiarso, W. (2015). Aplikasi Pemodelan Rasch: pada Assessment Pendidikan. *Trim Komunikata*, *September*, 1–24.

Suneja, S. (2023). Standardization of Objective Structured Practical Examination in Terms of Validity and Reliability in Biochemistry: Our First Experience. *Journal of Health and Research*, *10*(2), 167–171.

Tsai, M. J., Liang, J. C., & Hsu, C. Y. (2021). The Computational Thinking Scale for Computer Literacy Education. *Journal of Educational Computing Research*, *59*(4), 579–602.

Wati, F., Sinaga, P., & Priyandoko, D. (2017). Science Literacy: How do High School Students Solve PISA Test Items? *Journal of Physics: Conference Series*, *895*(1).

Wati, M., Mahtari, S., Muthi'ah, A., Dewantara, D., & Suharno, S. (2023). The Scientific Literacy Test Instrument on Particle Dynamics for High School Students. *Asian Journal of Assessment in Teaching and Learning*, *13*(2), 46–56.

Wright, B. D., & Masters, G. N. (1982). *Ratin Scale Analysis Rasch Measurement*. Mesa Press.

Yilmaz, E. (2022). Development of Mindset Theory Scale (Growth And Fixed Mindset): A Validity and Reliability Study (Turkish Version). *Research on Education and Psychology*, *6*(Special Issue), 1–26.

Yusmaita, E., Yuliani, F., & Gazali, F. (2022). Analysis of Chemical Literacy Instrument on Oxidation and Reduction Reactions Material By Using Rasch Model. *JCER (Journal of Chemistry Education Research)*, *6*(2), 124–130.

Yusup, M. (2021). Using Rasch Model for The Development and Validation of Energy Literacy Assessment Instrument for Prospective Physics Teachers. *Journal of Physics: Conference Series*, *1876*(1).

Zhang, L., Liu, X., & Feng, H. (2023). Development and Validation of An Instrument for Assessing Scientific Literacy From Junior to Senior High School. *Disciplinary and Interdisciplinary Science Education Research*, *5*(1).