**Construct Validity of the Test Instrument of Critical Thinking Skill on STEM-PjBL using Rasch Model**

Rina Dwi Setyawati1, a), Agnita Siska Pramasdyahsari 2, b), Zaenuri3,c),Iqbal Kharisudin4,d), Julia Puspita Arum5,e), Iin Dwi Astutik6,f), Sindi Nur Aini7,g), Ukima Nusuki8,h),Ummi Salmah9,i), Nafiatus Zuliah10,j), Yusuf Arkham Prihandika11,k), Budi Cahyono12,l), and Binod Prasad13,m)

Author Affiliations

1,2,5,6,7,8*Department of Mathematics Education, Universitas PGRI Semarang, Indonesia* 3,4,11*Department of Mathematics Education, Universitas Negeri Semarang, Indonesia*

12*Department of Mathematics Education, Universitas Islam Negeri Walisongo Semarang, Indonesia*

9 *Seameo Yogyakarta*

10*SMP Baruna Wati Semarang*

*13Kathmandu University, Nepal*

 *Author Emails*

a) Corresponding author: rinadwisetyawati@upgris.ac.id

b)agnitasiskapramasdyahsari@upgris.ac.id

c) zaenuri.mipa@mail.unnnes.ac.id

d) iqbalkharisudin@mail.unnes.ac.id

e) yusufarkhamp@students.unnes.ac.id

f) budi.cahyono@walisongo.ac.id

m)*binod@kusoed.edu.np*

**Abstract.** Students' critical thinking skills are very important abilities in the 21st century. To measure critical thinking skills, a valid and reliable instrument is needed. This article uses the Rasch model to construct an instrument for critical thinking skills on number pattern material whose learning uses STEM PjBL. The method used is quantitative research by measuring the validity and reliability of experimental instruments with Rasch analysis using the Winstep program. The instrument was made based on indicators of critical thinking ability. The instrument consists of 15 items with a description type that is validated by experts. The instrument was tested on 33 students at SMP Negeri 6 Semarang and on social arithmetic subjects. The results of the Rasch model analysis show that the questions have a moderate level of difficulty. At PT-Measure Corr. indicates that the discriminating power of the grains has a high discrepancy. Furthermore, checking whether an item is appropriate (item fit) or not fit (outlier or misfit) is shown in the MNSQ OUTFIT value, ZSTD OUTFIT value, and PT MEASURE CORR value. From the questions tested, those that meet the fit items are 10 questions from 15 questions. Cronbach's Alpha value (KR-20) is the reliability coefficient obtained based on the classical test theory approach. This value describes the interaction between a person and the item as a whole. The alpha value is 0.72, which indicates that the reliability of the test is generally very satisfactory (reliable). The value of person reliability is 0.76, and the value of item reliability is 0.84. This shows that the consistency of the answers from our subjects is good, and the quality of the items in the instrument's reliability aspect is good.

# Introduction

Critical thinking skills are indispensable in 21st century learning [1-2]. Critical thinking is an ability pattern-focused reflective thinking decision making about what must be believed, must be done and can proved correct. ([3-4] said that FRISCO's have six criteria are as follows (1) Focus means students answer questions in context problem, (2) Reason means that students can give reasons relating to facts or evidence relevant at every step make conclusions, (3) Inference means students can draw conclusions by appropriate based on the identification process on resolution steps, (4) Situation means students can collect relevant information and use relevant mathematical concepts to answer questions, (5) Clarity means students can provide clarity of symbols or things that are not clearly explained, (6) Overview means students have checked repeat the work from start to finish what generates the FRISCO criteria.In the 21st-century learning process, it is necessary to choose an innovative learning model. The Project Based Learning (PjBL) learning model based on STEM (Science Technology Engineering and Mathematics) is the right choice. STEM-PjBL is learning that uses a Project Based Learning model that is integrated with STEM and consists of 5 processes, namely: reflection, research, discovery, application and communication [5]. STEM focuses on teaching and training students to engage in critical thinking, inquiry, problem solving, collaboration, and engineering as design thinking [6]. STEM is an integrative approach that is able to facilitate students in developing skills in the 21st century [7-9].

PjBL-based STEAM learning has a significant impact on generating ideas, creative solutions and students' critical thinking [10]. The PjBL-STEM model affects the improvement of critical thinking skills [11]. STEAM learning can help students develop critical thinking skills, problem-solving skills and collaboration [12]. To measure the improvement of critical thinking skills, a good instrument is needed.

According to[13], conveying an instrument is generally defined as a tool that, because it meets academic requirements, can be used as a tool to measure an object or collect data on a variable. The research wants to measure critical thinking skills, so it is necessary to make a good test by looking at its validity and reliability. [14]states that a measurig device can be said to be a valid measuring device if the measuring device can measure what is intended to be measured precisely. In terms of validity and reliability, of course, it is influenced by (1) the instrument, (2) the subject being measured, and (3) the person taking the measurement. In terms of measurement, especially in education, of course, the most important thing is the correct measurement result information. The results of measurements that are not precise will provide inaccurate information, so the conclusions drawn are also inaccurate.

 [15] states that validity comes from the word validity which means the extent to which the accuracy and precision of a measuring instrument (test) in carrying out its measuring function. A test is said to have high validity if the tool performs the measuring function correctly or provides measurement results that are in accordance with the purpose of the measurement. Rasch measurement model can prove that an instrument has a high level of validity and reliability. This is because the use of the Rasch model is a solution to the validity problem where the Rasch model provides useful statistics and offers a tremendous opportunity to investigate validity [16]. In addition, the application of the Rasch model in a study will be able to facilitate and produce more efficient, reliable, and valid measurements in addition to increasing user convenience. The Rasch model can produce reliable and valid instruments [17][18]. A study to identify the validity and reliability of the instrument is very important to maintain the accuracy of the instrument [19].

This is necessary to ensure that the instrument can measure what it intends to measure consistently and accurately. The Rasch model considers the ability of the respondent to answer each item or question as well as the level of difficulty of the item itself [20]. The item fit analysis can evaluate whether the items in the instrument can measure what should be measured. On the other hand, if the item does not match (misfit), it is said that the item measures a construct outside the instrument, so revision or elimination of the Smith item in Napitulu must be carried out. Based on [21], criteria to determine the level of validity and reliability of the instrument. Validity is the extent to which a research test instrument measures what it is supposed to measure. Therefore, good conclusions can be made from the research sample[22].

## Method

The approach used in this research is a quantitative approach with the method of developing test instruments with descriptive analysis of the test results. Critical thinking skills test. tested on 33 students and has met the number of respondents who have met the requirements where based on. [23]the number of respondents for the pilot study is between 25 to 100 people. [24] suggests the minimum number of respondents is 30 people. The analysis method of research data used is the Rasch model with the Winstep Software version 3.92.1 developed by [25]. The data was collected by using a test that was constructed by the researcher which contained 15 questions. The data analysis technique used RASCH modeling analysis. This research identifies several things which include; 1) reliability, 2) validity, 3) analysis of the difficulty level of the questions, 4) the distribution of items, and 5) a description of the projection of respondents' answers in the scalogram. These five things are sufficient to describe the measuring power of the instrument on students' critical thinking skills. In addition, this analysis will show the reliability of respondents in taking tests which will later be used as data to measure students' critical thinking skills.

[26]suggest the following criteria to check whether an item is fit (item fit) or not fit (outlier or misfit). The criteria is to look at the MNSQ OUTFIT value greater than 0.5 and less than 1.5, getting closer to 1 (Ideal), the ZSTD OUTFIT value greater than -2.0 and smaller than +2.0, getting closer to 0 (Ideal), and the value of PT MEASURE CORR is greater than 0.4 and less than 0.85. The three criteria must be met by the item, so that the item is considered fit. In Perspective, Rasch sees that the item difficulty level and the testee's ability level are on the same scale, namely the logit scale (moving between -4 to 4). A rough idea of ​​the logit number is the Testee (ability level). This taste will show low ability (-4.00 to -2.00), medium ability (-1.99.00 to 1.99), and high ability (2.00 to 4.00). The items indicating the level of difficulty (measures) are the easy items are in the MEASURE value of -4.00 to -2.00), the medium items are in the MEASURE value of 1.99 to 1.99, and the difficult items are in the MEASURE value of 2.00 to 4.00. Item Difference Power/PT-measure (Point Measure/Total Item Correlation) is shown if a negative value is obtained, meaning that item is problematic. If the item is below 0.20 it needs further inspection/not good.

# Result and discussion

# the expert validation and indicators of critical thinking skills

Product validation was carried out by two experts. Validation of product data from the two experts obtained an average expert validation of 3.6 in the very good category or directly usable. Based on the instrument validation, several aspects were found in the valid category with several revisions, namely the general construction aspect from the validity aspect. While the aspects of the validity of the items in the category of minor revisions. However, even though it is said to be very valid and can be used without revision, the researcher still corrects the shortcomings that exist in the instrument and is adjusted to the comments and suggestions from the validator.[27] states that valid data can be used to measure what should be measured. The following are indicators of critical thinking skills that are used as the basis for making questions, which can be seen in table 1.

Table 1. Description of Critical thinking indicator

|  |  |
| --- | --- |
| **Critical Thinking Indicator (FRISCO)** | **Indicators needed in competence** |
| F (Focus) | Students understand the problem in a given problem. Added understanding of what information is known about the problem |
| R (Reason) | Students provide reasons based on relevant facts/evidence at each step in making decisions and conclusions. In this step, students should make steps to solve the problem |
| I (Inference) | (1) Students make appropriate completion plans; (2) Students choose the right reason (R) to support the completion plan that has been made. |
| S (Situation) | Students use all the information that is appropriate to the problem. |
| C (Clarity) | (1) Students use further explanations about what is in the conclusions made; (2) If there is an unusual (new) term in the question, students can explain it; (3) Students give examples of similar cases. |
| O (Overview) | Students research or re-examine thoroughly from start to finish (produced by FRISC) |

## The Difficulty Power and Discrimainating Power

The Measure shows that the item difficulty level and the testee's ability level are on the same scale, which is meeting the logit scale, the value is between -4 to 4. From the test questions, it shows a moderate level of difficulty. No one goes on a high or low level of difficulty. The usefulness of the item difficulty level has two things, namely usefulness for teachers, testing, and teaching [28]. Its uses for teachers are (1) as an introduction to the concept of relearning and providing input to students about learning outcomes, and (2) obtaining information about curriculum emphasis or reviewing the usual items [29]. The level of difficulty of the items can also be used to predict the measuring instrument itself (question) and the ability of students to understand the material taught by the teacher. For example, one item is included in the easy category, so the predictions for this information are (1) the distractor of the item does not work and (2) most of the students answer the item correctly; This means that most of the students have understood the material being asked. If an item is included in the difficult category, then the prediction of this information is (1) the item "may" have the wrong answer key; (2) the item has two or more correct answers; (3) the material in question has not been taught or the learning has not been completed; so that the minimum competencies that must be mastered by students have not been achieved; (4) the material measured is not appropriate, it is asked using the form of the question given; and (5) the statement or sentence of the question is too complex and long, Figure 1 shows that the level of difficulty of the test questions is in the medium category seen in “Measure”.

Figure 1. Rasch's perspective for the Differentiating Power of Items in the test questions



Figure 1 also shows Rasch's perspective for the Differentiating Power of Items in the test questions, all items have high discriminating power at PT. Measure Corr. get more than 0.2. The discriminatory power of a question is the ability of the question to distinguish between students who have a high level of ability and students who have low abilities. [30]stated that the main purpose of Item Response Theory is to provide similarities between the statistical questions and ability estimates. [25] said that the use of the Rasch model is highly recommended to develop instruments to measure critical thinking skills in order to get valid and reliable questions

Figure 2. Summary Statistic

|  |  |
| --- | --- |
|  |  |

From Figure 2 show that the value of Cronbach's Alpha (KR-20) is the reliability coefficient obtained based on the classical test theory approach. This value is the interaction between people and goods as a whole. The alpha value is 0.84. This shows that the reliability of the test in general is very satisfactory (Reliable). The person reliability value is 0.81 and the item reliability value is 0.94. This shows that the consistency of the answers from the subjects is good and the quality of the items on the reliability aspect of the instrument is good. Reliability is the extent to which a research test instrument can be expected to obtain consistent results when the test is repeated. Reliability can provide consistency of validity [31].

## Viewing the Item-Person Map

One of the features of Rasch's analysis with Winstep is that there is a map that depicts the distribution of the subject's ability and the distribution of item difficulty with the same scale. This map is called the Wright Map which is nothing but a person-item map. This shows that the average individual ability is slightly higher than the item difficulty level. Therefore, this study will look further at one or two items that have a low index of difficulty to be changed and adjusted so that the instrument items will be a little more difficult for research respondents to answer.

 Figure 3. Wright Map



The distribution of respondents' abilities and the distribution of the level of difficulty of the questions with the same scale can be seen from the Wright Map. The Wright map of the results of the instrument trial can be seen in Figure 2. The Wright map on the left describes the students' abilities, and it can be seen that students with code P17 (have the highest ability compared to other students. Even though they have the highest ability, this student's logit score is less of +1 logit Student P17 is outside the limit of two standard deviations (T) indicating a different high intelligence (outlier). The student with the lowest ability is P11, with a logit value of more than -1, which also indicates a very low ability (outlier), because it is outside the boundary of T.

Wright's map on the right explains the distribution of the logit value of the item, number 4 (4d) is the problem with the highest level of difficulty (+1 logit), which means the probability that all students do this problem correctly. very small. As for question 1a is the question with the lowest logit value (close to -1 logit). In this case, the more students are able to work on this question I correctly. Problems 1b and 1d have the same level of difficulty because the logit value is the same, as in questions 1c and 2b. The average logit value of the item is always set in logit 0.0 which indicates the initial reference point of the scale. From the Wright Map, the average logit person was found to be -1 logit (below 0.0 logit). This shows that the average ability of students is below the average level of difficulty of standard questions. Wright's map also shows that 2 students had logit scores below -1 logit, reflecting that more than half of the sample had low critical thinking skills. This should be a concern when critical thinking skills are low because critical thinking skills are important when children learn because they are used to support when solving problems.[32]

# CONCLUSION

Rasch model analysis on the measure that shows the questions have a moderate level of difficulty. At PT-Measure Corr. indicates that the discriminating power of the grains has a high discrepancy. Furthermore, checking whether an item is appropriate (item fit) or not fit (outlier or misfit) is shown in the MNSQ OUTFIT value, ZSTD OUTFIT value, and PT MEASURE CORR valueFrom the questions tested, those that meet the fit items are 10 questions from 15 questions. Cronbach's Alpha value (KR-20) is the reliability coefficient obtained based on the classical test theory approach. This value describes the interaction between a person and the item as a whole. The alpha value is 0.72, which indicates that the reliability of the test is generally very satisfactory (reliable). The value of person reliability is 0.76, and the value of item reliability is 0.84. This shows that the consistency of the answers from our subjects is good, and the quality of the items in the instrument's reliability aspect is good.od

# Acknowledgments

Thank you to KEMENTERIAN PENDIDIKAN, KEBUDAYAAN RISET DAN TEKNOLOGI that has provided funding

# References

1. S. Lamb S, Q. Maire, and E. Doecke, *Key Skills for the 21st Century: an evidence-based review* (Education: Future Frontiers, State of New South Wales, 2017)
2. B. Cahyono, K. Kartono, B. Waluya, M. Mulyono, and R. D. Setyawati, *Cypriot Journal of Educational Sciences*, *16*(6), pp. 2956-2969 (2021)
3. R. H. Ennis, *Inquiry: Critical Thinking across the Disciplines* 26 (1): PP. 4–18 (2011)
4. B. Cahyono, K. Kartono, B. Waluyo, and M. Mulyono*, Journal of Physics: Conference Series*, 1321 (2) pp. 1742-6596 (2019)
5. D. Laboy-Rush, *Integrated STEM Education through Project-Based Learning*, (Learning.Com, 2011)
6. S. Han, R. Capraro and M. Capraro, *International Journal of Science and Mathematics Education* 13 (5) pp. 1089-1113 (2015)
7. E. Yulianti, E. Hamimi, and R. Rahmagusviana, *Phenomenon: Jurnal Pendidikan MIPA*, 11 (2), pp. 169-188 (2021)
8. G. Sarican, I. Technology, and D. Akgunduz, *Cypriot Journal of Educational Sciences* , 13 (1), pp. 94-107 (2018)
9. E. Saxton, R. Burns, S. Holveck, S. Kelley, D. Prince, N. Rigelman, and E. A. Skinner, *Studies in Educational Evaluation*, 40, pp. 18–35 (2014).
10. A. Fitriyah, and S. D. Ramadani, *Journal Of Chemistry And Education (JCAE)*, *X*(1), pp. 209–226 (2021)
11. T.R. Allanta, and L. Puspita, *Jurnal Inovasi Pendidikan IPA*, *7*(2), pp. 158–170 (2021)
12. S.N. Mufida, D. V. Sigit, and R.H. Ristanto, R. H, *Biosfer*, *13*(2), PP. 183–200 (2020)
13. Djaali, P. Muljono and Ramly, *Pengukuran Dalam Pendidikan* (Grasindo, Jakarta, 2008)
14. W. Nurkancana, and PPN. Sunartana, *Evaluasi Hasil Belajar* (Usaha Nasional, Surabaya, 1992)
15. R.D. Setyawati, *Phenomenon: Jurnal Pendidikan MIPA*, *7*(2), pp. 174-186 (2018)
16. S. Azwar, *Sikap Manusia Teori dan Pengukurannya* (Liberty, Yogyakarta, 1987)
17. T.G. Bond, and C.M. Fox, *Applying the Rasch Model : Fundamental Measurement in the Human Sciences Second Edition University of Toledo.an Sciences Second Edition,* LAWRENCE ERLBAUM ASSOCIATES, Mahwah, New Jersey London (2007)
18. A. Aziz, M. S. Jusoh, M. H. Amlus, A. R. Omar, and T.S. A. Salleh, *Journal of Applied Science and Agriculture*, *9*(12), pp. 7–12 (2014)
19. S. Ariffin, B. Omar, A. Isa, and S. Sharif, *Procedia - Social and Behavioral Sciences*, *9*, pp. 729–733 [(2010)](https://doi.org/10.1016/j.sbspro.2010.12.225)
20. G. Rasch, G. *Danish Yearbook of Philosophy*, Copenhagen: Hunksgaard 14 (2), pp. 68-94 (1980)
21. B. D. Wright, and M. H. Stone, *Test Design*. Mesa Press, Chicago (1979)
22. J. W. Creswell, *Educational research : planning, conducting, and evaluating quantitative and qualitative research- 4th ed*. (Pearson Education, Boylston Street, Boston, 2012)
23. D.R. Cooper, and P.R. Schindler, *Business research methods* (8th ed.). (McGraw-Hill, New York, 2003)
24. G.A. Johanson, and G.P. Brooks, *Educational and psychological measurement*, *70* (3), pp. 394-400 (2010)
25. J. M. Linacre, *Journal of Applied Measurement*, *7*(1), pp. 129-139 (2006)
26. W.J. Boone, *CBE Life Sciences Education*, *15*(4), Winter rm4 (2016)
27. P. D. Sugiono, *Metode Penelitian Kuantitatif dan Kualitatif* (Alfabeta, Bandung, 2015)
28. Safari, *Teknik analisis butir soal instrumen tes dan nontes* (assosiasi Pengawas Sekolah Indonesia-Depdiknas, 2015)
29. A.A. Rangkuti, *Perspektif Ilmu Pendidikan*, *23*(XIV), pp. 66-73 (2011).
30. A.W. Sabekti, and F. Khoirunnisa, *Jurnal Zarah*, *6*(2), pp. 68–75 (2018).
31. R. M. Yasin, F. A. N. Yunus, R. C. Rus, A. Ahmad, and M. B. Rahim, *Procedia-Social and Behavioral Sciences*, *204*, pp. 212-217 (2015)
32. Ruamba, M. Y., Mastur, Z., & Kharisudin, I. (2022). Kemampuan Representasi Matematis Calon Guru Matematika pada Mata Kuliah Teori Peluang. *Jurnal Pendidikan Matematika (JUPITEK)*, *5*(1), 34-41.