



Unraveling Scientific Creativity in Fluid Dynamics: An Item Response Theory Approach with Rasch Analysis for High School Learners

Maimon Sumo^{1*}, Finsensius Oetpah², Evangeline D. Pagal³

¹Universitas Islam Madura, Indonesia

²University of Timor, Indonesia

³Davao Del Sur State College, Digos City, Philippines

Abstract

Scientific creativity is a key 21st-century skill that should be assessed to identify students' creative potential. This is important for mapping out and guiding students toward the fields of interest they will pursue in the future. However, the measurement of scientific creativity in schools is still dominated by conventional ordinal score-based approaches, which fail to clearly reflect individual students' abilities. This indicates a gap between modern assessment expectations and limited evaluation practices. This study aimed to develop the Scientific Creativity Test in Fluid Dynamics (SCTDF) for high-school students and to conduct a preliminary psychometric evaluation of its measurement functioning using the Rasch model. This study used the 4D Research and Development model and involved 50 high school students in Pamekasan Regency, East Java. The research instrument was a test consisting of 14 multiple-choice items based on seven indicators of scientific creativity and fluid dynamics. Data analysis was conducted using the Ministep application according to the Rasch Model criteria. The findings provide preliminary Rasch-based evidence regarding the measurement functioning of the SCTDF. Most items demonstrated acceptable model-data fit; however, one item showed misfit and person reliability was low (0.49), indicating limited precision in differentiating respondents along the measured continuum. Further item refinement and validation with larger and more diverse samples are therefore required. The study concludes that the SCTDF can serve as an alternative tool for assessing scientific creativity based on modern psychometrics, although revisions to certain items are still required to improve measurement accuracy.

Keywords: Scientific creativity; fluid dynamic; item response theory; Rasch model

(*) Corresponding Author: maimonsumo@uim.ac.id.

How to Cite: Sumo, M., Oetpah, F., & Pagal, E. D. (2026). Unraveling Scientific Creativity in Fluid Dynamics: An Item Response Theory Approach with Rasch Analysis for High School Learners. *Formatif : Jurnal Ilmiah Pendidikan MIPA*, 16(2), 397-410. <https://doi.org/10.30998/formatif.v16i2.3841>

INTRODUCTION

The 21st-century education paradigm has undergone a very rapid transformation. The conventional concept of education as a process of knowledge transfer has shifted towards fulfilling complex skills needs. Among the skills required in the 21st century are critical thinking and problem-solving, creativity and innovation, communication, and collaboration, known as 4C skills (Sumo & Soeharto, 2025). Creativity, a 21st-century skill, is known in science as scientific creativity. This skill is essential for every student to master in facing the challenges of an increasingly advanced era, both in terms of knowledge and technology (Dilekçi & Karatay, 2023). Scientific creativity can help students solve

problems logically and innovatively (Pinar et al., 2025). Fluid dynamics is a physics topic considered to foster students' scientific creativity (Hikmah et al., 2025).

Research by Prahani et al. (2021) assessed students' scientific creativity using an ordinal scale: scores of 2 for <5%, 1 for 5–10%, and 0 for >10%. However, the study did not clearly evaluate the test's validity and reliability. Another study by Irma et al. (2023) assessed creativity based on response percentages for each indicator, but did not align the creativity of individual students with the test items. In Liu et al. (2023), students' creativity was scored as 0 or 1, leaving the creativity assessment unclear. The study by Xu et al. (2024) used an ordinal scale to measure creativity and considered gender differences, but did not provide details on test items suitable for each gender.

Despite the availability of scientific-creativity assessments, uncertainty remains regarding how domain-specific scientific-creativity tasks function as a measurement continuum, whether different proposed creativity facets are adequately represented, and whether item functioning remains comparable across student subgroups. Fluid dynamics provides a theoretically relevant context because solving and designing fluid-related problems requires the coordination of scientific knowledge with problem identification, evaluation, imagination, and design. The study tested the measurement functioning of domain-specific scientific-creativity tasks, not simply introducing Rasch. That reason constitutes the research gap for this study.

The contribution of this study lies not in introducing Rasch analysis as a new technique, but in examining how seven proposed facets of scientific creativity are represented within fluid-dynamics tasks, identifying the empirical hierarchy and measurement functioning of these tasks, and identifying items requiring refinement in a domain-specific physics assessment (Firdaus et al., 2025). It identifies which items are easy or difficult for students to answer and checks if items fit the scientific creativity construct (validity item fit) (Shaw et al., 2021). It assesses the instrument's consistency and the stability of students' responses (instrument reliability). It determines if items favor males or females (gender bias/differential Item Functioning/DIF) (Panadero et al., 2023). It maps students' abilities and item difficulties on the same scale (Wright Map) and checks if the rubric is effective (Uba & Khairani, 2024).

In this study, the scientific creativity test dynamic Fluid/SCTDF was developed not only to measure students' scientific creativity test results but also to evaluate the test's difficulty, individual ability to answer, item validity, item reliability, item difficulty, and item bias related to gender and students' origin. Thus, the SCTDF created in this study can assess students' scientific creativity levels and serve as a reference for physics tests, especially on dynamic fluids. This study is guided by two previously formulated research questions:

RQ1: To what extent can research data collected from samples of high school students and analyzed using the Rasch model measure scientific creativity?

RQ2: Are these scientific creativity questions free from gender and student background bias so that they can measure students' abilities fairly?

METHODS

Grade Science-track students from public high schools in Pamekasan Regency, Madura, East Java, Indonesia participated in this study. The sample consisted of 50 students, comprising 25 males (12 from rural areas and 13 from urban areas) and 25 females (12 from rural areas and 13 from urban areas). The sample was selected using purposive

sampling; this method was chosen based on the research objectives, and because the researcher could not reach all respondents but only represent each of the research objectives (Creswell, 2014). Given the small subgroup sizes, the DIF analyses were treated as exploratory and hypothesis-generating rather than confirmatory evidence of measurement invariance.

This research employs 4D research and development model; this type of research was chosen with the aim of producing a scientific creativity test product that is valid, practical, and effective in generating a reliable product (Zaky et al., 2021). The first stage is Define, which involves a literature review on scientific creativity, an analysis of the fluid dynamics curriculum and scope, and the establishment of the construct domain. The second stage is Design, involving the creation of a test blueprint that links seven indicators, fluid dynamics concepts, cognitive demands, and item numbers. The third stage is Develop, consisting of test trials to determine the validity and usability of each item (Leung et al., 2018); the fourth stage is Disseminate, involving implementation as a pilot administration.

The multiple-choice format was selected to support standardized administration, objective scoring, and dichotomous Rasch calibration. The items were designed to require students to evaluate scientifically plausible alternatives within fluid-dynamics contexts rather than merely recall factual knowledge. Nevertheless, because the selected-response format constrains the generation of multiple original responses, the SCTDF should primarily be interpreted as assessing constrained scientific-creative reasoning rather than the full spectrum of divergent scientific creativity (Liu et al., 2024). Each SCTDF item is accompanied by answer choices from A to E. Each SCTDF item is scored “1” for a correct answer, and “0” for an incorrect answer (Firdaus et al., 2025; Hu & Adey, 2002). The seven indicators serve as content facets of a provisional overarching construct rather than as seven separate subscales. While two items per facet were included to ensure balanced content sampling in the initial version of the instrument, reliability scores or interpretations should not be derived from just two items per indicator. A test blueprint was employed to examine dimensionality specifically through Rasch PCA of residuals and local dependence. Research by Xu et al. (2025) also demonstrates that scientific creativity possesses a complex structure; therefore, dimensionality should not be assumed based solely on indicator labels. For details on the test based on scientific creativity indicators and intended cognitive operations, see Table 1.

Table 1. Scientific creativity test blueprint indicators

Scientific-creativity facet	Operational definition	Fluid-dynamics concept/context	Items	Intended cognitive operation
Unusual use	Generating many original ideas	Debit	Q1	Connect scientific concepts or ideas
		Principle of continuity	Q8	
Problem finding	Finding a different idea	Debit	Q2	Problem identification
		Principle of continuity	Q9	
Scientific imagination	Generating logical ideas	Debit	Q3	Mentally constructing
		Principle of continuity	Q10	
Product improvement	Improving product quality	Debit	Q4	Divergent thinking
		Principle of continuity	Q11	
		Debit	Q5	

Scientific-creativity facet	Operational definition	Fluid-dynamics concept/context	Items	Intended cognitive operation
Creative science problem solving	Solve problems creatively	Principle of continuity	Q12	Divergent scientific problem solving
Creative experiment designing	Designing creative experiments	Debit Principle of continuity	Q7 Q13	Original synthesis
Creative product design	Designing creative products	Debit Principle of continuity	Q6 Q14	Synthesize scientific knowledge

The data were analyzed in line with the research goals and questions, focusing on validity, reliability, test item fit, and Differential Item Functioning (DIF) (Ranganathan et al., 2024). Item fit was checked to ensure the items measured what they were supposed to. The validity of SCTDF items was evaluated using these criteria: outfit mean-square (MNSQ) between 0.5 and 1.5, Outfit Z-standard (ZSTD) between -2.0 and +2.0, and Point-Total Measure Correlation (PTMEA) between -0.4 and 0.85 (Zhao et al., 2025). The reliability of the SCTDF items was assessed to see if the test consistently measured students' scientific creativity. SCTDF reliability includes item, person, and item-person (Cronbach, 1951). Cronbach's alpha coefficient ranges from 0 to 1, with scores near +1 indicating high reliability. After calculating the Cronbach's alpha score, it was interpreted using Table 2.

Table 2. Interpretation of reliability scores based on Cronbach's alpha coefficient

Statistical summary	Coefficient value	Interpretation
Item and test-taker reliability	$r > 0.94$	Excellence
	$0.90 < r \leq 0.94$	Very high
	$0.80 < r \leq 0.90$	high
	$0.67 < r \leq 0.80$	Currently
Cronbach' s alpha coefficient (KR-20)	$r \leq 0.67$	Low
	≥ 0.9	Very high
	$0.6 \leq \alpha < 0.9$	Currently
	$0.5 \leq \alpha < 0.6$	Less
	$\alpha \leq 0.5$	Low

In the Rasch analysis, the probability of each test item being answered correctly by a person was calculated. They were then adjusted based on how well the items matched the test-takers (Tutz, 2025). This helped identify items that did not match well (outliers) (Chan et al., 2021). The probability shows how test-takers relate to items. They also checked if items were biased based on gender and location using Differential Item Functioning (DIF). DIF helps identify bias in test items related to gender and other factors (Kiefer & Sengewald, 2025). Scientific creativity tests are biased if the DIF t-value is below -0.2 or above +2.0, the DIF contrast value is less than -0.5 or more than +0.5, and the p-value is less than 0.05 or more than -0.05 (Li et al., 2025). The DPF curve shows how well the test items measured students' scientific creativity. It checks whether the test items work the same for all gender groups (Engelhard & Wind, 2026). ICC shows the characteristics of the test items. The red curve represents the standard curve of the test and the blue curve represents the observed curve (Akbari, 2025). The IIF graph helps identify items as moderate, easy, or difficult (Iriyadi et al., 2024).

RESULTS & DISCUSSION

Results

The study results are presented as quantitative descriptions using the Rasch model with the help of the Ministep tool. The findings answer the research questions and meet the research objectives. The study provided validity data based on infit and outfit MNSQ and ZSTD, and reliability data for individuals, items, and both together, shown by Cronbach's alpha (KR-20). For detailed results, see Table 3.

Table 3. Summary Statistics of SCTDF test persons and items

Person	Outfit	
	MNSQ	ZSTD
Mean	1.09	-0.02
Person Reliability = 0.49		
Cronbach Alpha (KR-20) = 0.78		
Item		
Mean	1.09	-0.02
Item Reliability = 0.85		

The test's validity, assessed from both item and person perspectives as shown in Table 2, is considered good. This is shown by the mean MNSQ score for individual validity of 1.09, which meets the MNSQ acceptance criteria. The test item validity is 0.09, and the average standard deviation/ZSTD is -0.02, which is above -0.5. The person reliability of 0.49 indicates limited precision in separating respondents along the continuum measured by the current 14-item SCTDF. Accordingly, individual-level classification is not supported by the present pilot data. For more details on the validity of each test item for each scientific creativity indicator, see Table 4.

Table 4. Rasch item-fit statistics for SCTDF items

Item	Outfit MNSQ	Outfit ZSTD	PTMEA	Category
Q1: <i>Unusual use</i>	1.49	1.16	0.07	Acceptable fit
Q2: <i>Problem finding</i>	1.72	1.24	0.05	Misfit
Q3: <i>Scientific imagination</i>	1.30	0.95	0.22	Acceptable fit
Q4: <i>Product improvement</i>	0.94	-0.12	0.44	Acceptable fit
Q5: <i>Creative science problem solving</i>	0.99	-0.01	0.43	Acceptable fit
Q6: <i>Creative product design</i>	0.77	-1.43	0.57	Acceptable fit
Q7: <i>Creative experiment designing</i>	0.91	-0.47	0.52	Acceptable fit
Q8: <i>Unusual use</i>	1.20	0.56	0.26	Acceptable fit
Q9: <i>Problem finding</i>	1.14	0.47	0.29	Acceptable fit
Q10: <i>Scientific imagination</i>	1.21	0.89	0.24	Acceptable fit
Q11: <i>Product improvement</i>	0.93	-0.27	0.59	Acceptable fit
Q12: <i>Creative science problem solving</i>	1.03	0.23	0.44	Acceptable fit
Q13: <i>Creative experiment designing</i>	0.98	-0.08	0.45	Acceptable fit
Q14: <i>Creative product design</i>	0.62	-1.81	0.66	Acceptable fit

Thirteen of the fourteen items demonstrated acceptable model–data fit under the prespecified fit criteria, whereas Q2 showed misfit. These statistics are interpreted as evidence of item model compatibility and not as stand-alone evidence of construct validity, exceeding the limit of 1.5. The ZSTD values for all SCTDF items were between -2.0 and +2.0, which is within the Rasch model's range. Test items based on scientific creativity indicators showed acceptable variance for the Rasch model. Another important factor is

PTMEA. The Rasch model analysis indicated that the scientific creativity test items met the PTMEA criteria. The SCTDF scientific creativity items ranged from 0.07 to 0.66, according to the PTMEA criteria. The positive PTMEA values indicate that item responses were generally aligned in the expected direction with the overall measured variable. This evidence is interpreted together with, rather than as a substitute for, other sources of validity evidence. Figure 2 illustrates how well the items fit the data.

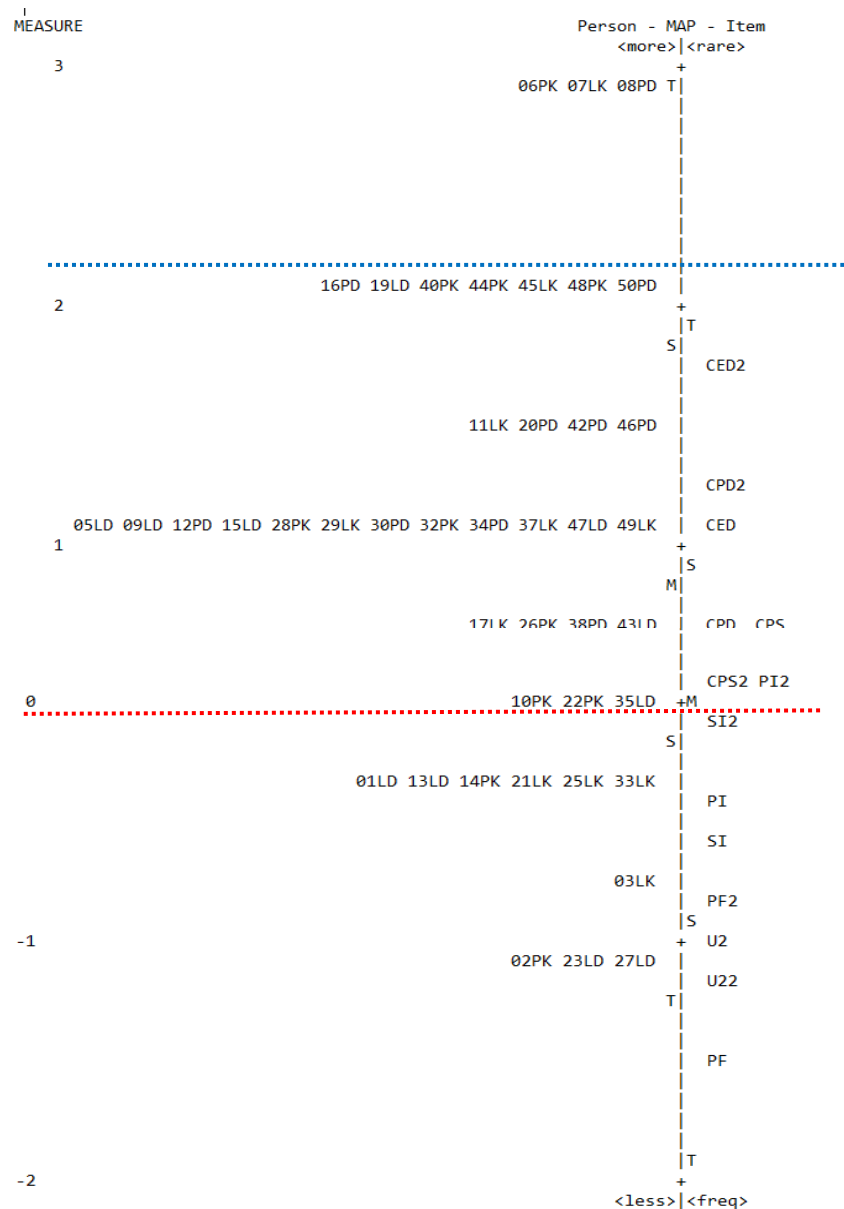


Figure 1. Variable Wright Maps

Figure 1 shows the logit scale describing students' abilities to answer the test questions. A higher logit score indicates that the question is more difficult and suitable for students with higher abilities. In the dichotomous Rasch model, the probability of a correct

response is 0.50 when person ability (θ) equals item difficulty (b). Therefore, a person measure of 0 logits implies a 50% probability only for an item located at 0 logits, not for all items. The statement can be proven using the equation $P(X = 1) = e^{(\theta-b)} / [1 + e^{(\theta-b)}]$. The map above shows that students with the codes 10 PK, 22 PK, and 35 LD obtained scores of 7 out of 14.

Students' test performance is divided into three groups: higher students can tackle difficult tests like 17 LK, 26 PK, 38 PD, 43 LD, 16 PD, 19 LD, 40 PK, 44 PK, 45 LK, 48 PK, 50 PD, 05 PK, 07 LK, and 08 PD; medium students can handle tests like 10 PK, 22 PK, 35 LD, 17 LK, 26 PK, 38 PD, and 43 LD; and lower students can manage tests like 02 PK, 23 LD, 27 LD, 03 LK, 01 LD, 13 LD, 14 PK, 21 LK, 25 LK, and 35 LK. Lower-ability students can answer questions on scientific creativity, such as unusual uses or explaining pipe size in fluids. High-ability students deal with complex tasks, like designing experiments and identifying variables. After confirming the questions are suitable, the next step is to check for question bias using DIF and DPF. To identify item bias related to gender and students' origin, we refer to the results of the Differential Item Functioning (DIF) analysis. DIF can be used to examine test item bias with respect to a given difference, as shown in Figure 2.

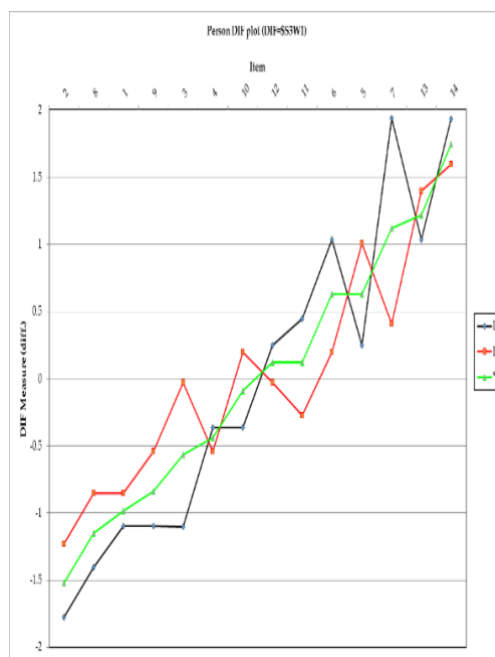


Figure 2. Person DIF plot (Male and Female)

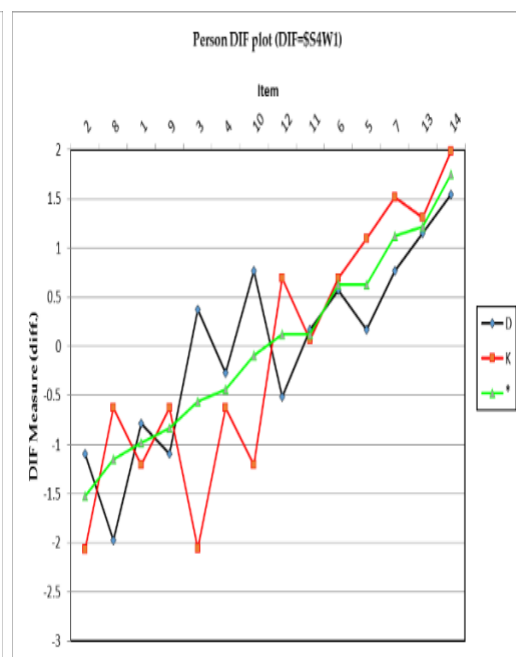


Figure 3. Person DIF Plot (City and Village)

According to the DIF Plot analysis, curves nearing the top, like item 7, show that this item is easier for male students and harder for female students. In item 7, students must design an experimental model of dynamic fluids, focusing on water flow with different cross-sectional areas, considering water pressure and velocity. Item 3 was easier for female students; here, students needed to express their ideas on improving product quality. The graph indicates that male students excel in items 6, 7, and 14, while female students perform better on items 3, 5, and 10. The curves for males and females do not overlap, indicating no biased test items. However, item 7 has a t -value of +1.7, below +2.0, and a probability value of 0.0312, under 5%. By these standards, the test item is not biased. To detect item

bias based on students' residence or origin, refer to the DIF analysis results shown in the DIF graph in Figure 3.

According to the DIF Plot analysis (Figure 3), the red graph shows city students (K) who could answer test items 5, 6, 7, 11, 12, 13, and 14 with a maximum logit scale of +1.98. The black line shows village students (D) who could answer test items 3, 5, 6, 7, 10, 11, 13, and 14 with a maximum logit scale of +1.54. Village students answered more test items with positive logits, totaling eight items (average logit +0.64), while city students answered seven items (average logit +1.13). This indicates that village students excelled in the number of items answered, but city students showed higher ability, even with fewer items. For village students, biased test items based on the t value are items 3, 5, 7, 8, 10, and 12. Based on the contrast value, biased items are 2, 3, 5, 7, 8, 10, and 12, with the probability for both groups being ($p > 0.05$). The consistency of test items in measuring students' scientific creativity levels is shown in the DPF Plot curve in Figure 4. The graph shows that they slightly deviated from the ICC but were still acceptable. Overall, the test items were divided into easy, moderate, and difficult categories, as shown in Figure 5.

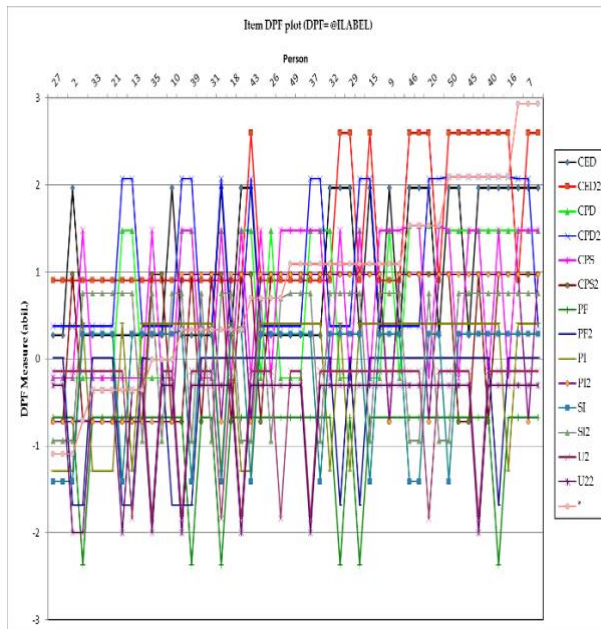


Figure 4. Item DPF Plot

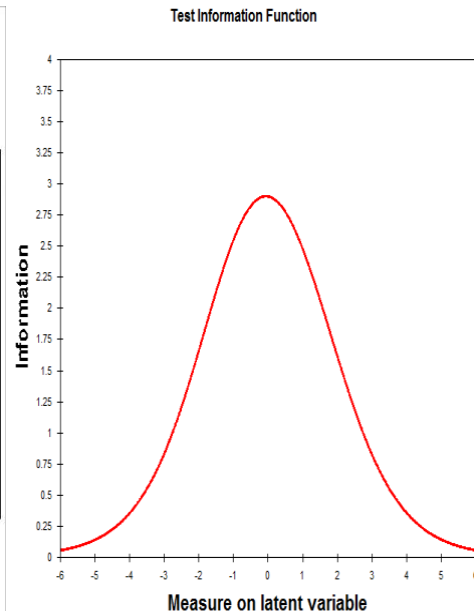


Figure 5. Item information function (IIF)

The DPF graph in Figure 5 helps determine whether a test item works the same for different groups. The information function indicates where along the latent continuum the SCTDF provides the greatest measurement precision. The DPF graph in Figure 5 shows that the test item "CED 2" measures scientific creativity in physics well for high-ability students, both male and female. The item "PF" measures low-level scientific creativity, shown by a negative score, for both male and female students from rural and urban areas. Test items that match students' abilities can be seen from the expected-score ICC. Figure 6 provides information on students' abilities to complete the scientific creativity test on the topic of dynamic fluids based on the test item results. From the graph, students' abilities are at a moderate level, with an information value greater than +2.5. This result indicates that

the scientific creativity test for physics on dynamic fluids is suitable for measuring students' abilities in the moderate category.

Discussion

RQ1: To what extent can research data collected from a sample of high school students and analyzed using the Rasch model measure scientific creativity?

The emerging item hierarchy suggests that the seven scientific-creativity facets may not impose equivalent cognitive demands when embedded in fluid dynamics. Tasks involving experimental design, for example, may require students to coordinate conceptual knowledge, variable identification, prediction, and evaluation simultaneously, whereas more constrained tasks may require fewer interacting cognitive operations (Hope et al., 2025). Valid test items can be applied in future measurement scales (Tsai et al., 2025). Besides checking the MNSQ mean score, each indicator's results can be examined. Rasch model analysis revealed that all indicator test items were valid, except for test item 2, which was not. Invalid items can be improved by studying or revising them. (Yuanyuan et al., 2025) demonstrated that revising test items helps maintain their accuracy or validity.

The SCTDF test showed reliability and consistency, with a Cronbach's alpha (KR-20) value of 0.78, indicating moderate reliability. Some test items need revision before using this creativity test (Aleksic et al., 2024). However, this test has not fully assessed students' scientific creativity. The coefficient of variation for student performance was 2.36, showing varied creativity levels among students. The individual logit mean was +0.74, higher than the item logit mean of 0.37, indicating good student performance on the test. A reliability standard deviation of +0.82 suggests that students' abilities are evenly distributed in solving scientific creativity questions (Firdaus et al., 2025). Test items that measure what they are intended to measure are valid and reliable (Engelhard & Wind, 2026).

Figure 2 illustrates how well students' abilities aligned with the test question difficulties. This alignment helps us gauge students' scientific creativity levels (Irey et al., 2026). The logit scale on the left indicates item difficulty and student ability. A positive logit value means the question is harder and needs more ability (Gittel et al., 2024). In Figure 2, twenty-six students had person measures above [criterion/reference point], indicating relatively higher performance on the SCTDF within this sample. Given the preliminary construct evidence and low person reliability, these measures should not be interpreted as categorical evidence of high scientific creativity. This suggests that high school students in Pamekasan from these two schools exhibit significant scientific creativity. The test questions on dynamic fluid mechanics were ordered from easy to hard. However, some students answered harder questions correctly but missed easier ones.

RQ2: Are scientific creativity questions free from gender and student-origin bias, thereby allowing for a fair assessment of students' abilities?

From the DIF graph, we observe that male students scored higher on test items 6, 7, and 14, while female students scored higher on items 3, 5, and 10. The graph shows no overlap between the male and female curves, suggesting no biased test items. However, Gender-related DIF evidence, particularly for Item 7, was inconclusive. Given the small subgroup sizes and inconsistent statistical and effect-size signals, these findings are hypothesis-generating and should not be interpreted as evidence that the SCTDF is free from gender-related measurement bias. This makes the item unclear or inconclusive

because it partly meets the DIF bias but is not supported by the graph (Pelouto et al., 2026). The DIF assessment uses criteria like the t-value, contrast value, and probability value, which should not be less than 5% (Touzani et al., 2024). This aligns with the research by Julianto et al., (2025) who stated that items not meeting the Rasch model need revision. (Yuanyuan et al., 2025) also showed that improving test items helps maintain their accuracy.

The scientific creativity test items aligned with the Rasch model, allowing them to assess high school students' scientific creativity. Thirteen of the fourteen items demonstrated acceptable model–data fit under the adopted criteria, whereas Q2 exceeded the Outfit MNSQ criterion and was flagged for further investigation and revision. Items U2, PF, and SI2 were easy for most students, with infit MNSQ values of 1.28, 1.20, and 1.19, respectively, which were higher than the average of 1.16. Items PI, CPD, CPS, CED, CPD2, CED2, PI2, U22, and CPS2 were less challenging. They showed minor deviations on a graph but remained acceptable (Zhao et al., 2025). According to the Rasch model, these items require low ability (Jin & Eckes, 2025). Their infit MNSQ values were above +0.5, consistent with the outfit MNSQ scores above 0.5.

The scientific creativity test items are effective and cover various aspects of students' creativity in physics, particularly in fluid dynamics. The ICC curve illustrates how well the items assess a person's ability, but it sometimes does not align with the actual ability (Lu et al., 2025). Test items that do not meet the criteria need revision. The more an item is revised, the more accurate it becomes (Hope et al. 2025). Another concern is measurement bias (Kangwanrattanakul & Kulthanachairojana, 2024). Individual responses offer valuable insights into the normality of items (Hong et al., 2021). Additionally, the IIF graph data indicated that the test items align well with Rasch model criteria (Iriyadi et al., 2024).

CONCLUSION

This study provides preliminary psychometric evidence of the functioning of the SCTDF among 50 high-school students. Thirteen of the fourteen items demonstrated acceptable Rasch model–data fit, whereas item Q2 showed misfit and requires substantive revision. The KR-20 coefficient (0.78) indicated moderate internal consistency and item reliability (0.85) suggested a relatively stable item hierarchy; however, person reliability was low (0.49), indicating limited precision in differentiating respondents along the measured continuum. DIF findings should be considered as exploratory because of the small subgroup sizes and unresolved evidence for particular items, including Item 7. Accordingly, the present findings do not establish the SCTDF as a fully validated measure of scientific creativity. Instead, they provide preliminary evidence to inform item refinement, dimensionality evaluation, and subsequent validation using larger and more diverse samples.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

ACKNOWLEDGEMENT

I want to thank Madura Islamic University for supporting this research. I also thank my colleagues for their important help in getting this research published.

REFERENCES

- Akbari, A. (2025). The rasch analysis of item response theory: an untouched area in evaluating student academic translations. *SKASE Journal of Translation and Interpretation*, 18(1). <https://doi.org/10.33542/JTI2025-1-05>
- Aleksic, V., Milic, M., Dotlic, J., Jeremic, B., Djerkovic, B., Radic, I., Karanovic, A., Kulic, L., Galjak, M., & Gazibara, T. (2024). The night eating questionnaire through the lens of the Rasch model. *Heliyon*, 10(11), e31929. <https://doi.org/10.1016/j.heliyon.2024.e31929>
- Chan, S.-W., Looi, C.-K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: a Rasch model measurement analysis. *Journal of Computers in Education*, 8(2), 213–236. <https://doi.org/10.1007/s40692-020-00177-2>
- Creswell, J. W. (2014). *Research Design: Pendekatan, Kualitatif, Kuantitatif, dan Mixed*. Pustaka pelajar.
- Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika*, 16(3), 297–334. <https://doi.org/10.1007/BF02310555>
- Dilekçi, A., & Karatay, H. (2023). The effects of the 21st century skills curriculum on the development of students' creative thinking skills. *Thinking Skills and Creativity*, 47(Marc 2023), 101229. <https://doi.org/https://doi.org/10.1016/j.tsc.2022.101229>
- Engelhard, G., & Wind, S. A. (2026). Evaluating measurement quality for dichotomous Rasch models. In *Invariant Measurement with Explanatory Rasch Models in the Human Sciences* (pp. 63–92). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-08816-1_4
- Firdaus, F., Wiyanto, W., Putra, N. M. D., & Isnaeni, W. (2025). Design of instruments for scientific creative thinking skills and creative thinking digital skills: Rasch models and confirmatory factor analysis. *Eurasia Journal of Mathematics, Science and Technology Education*, 21(5), em2632. <https://doi.org/10.29333/ejmste/16310>
- Gittel, L., Li, J. Q., Dell, J., Wintergerst, M. W. M., Heinz, C., Finger, R. P., & Terheyden, J. H. (2024). Piloting Rasch model scoring of the National Eye Institute Visual Function Questionnaire in uveitis. *Journal of Ophthalmic Inflammation and Infection*, 14(1), 16. <https://doi.org/10.1186/s12348-024-00398-x>
- Hikmah, N., Herayanti, L., & Gummah, S. (2025). Development of ethnoscience-based teaching materials on fluid dynamics to enhance students' creativity. *Jurnal Pendidikan Fisika Dan Teknologi (JPFT)*, 11(1). <https://doi.org/https://dx.doi.org/10.29303/jpft.v11i1a.9575>
- Hong, M., Lin, L., & Cheng, Y. (2021). Asymptotically corrected person fit statistics for multidimensional constructs with simple structure and mixed item types. *Psychometrika*, 86(2), 464–488. <https://doi.org/10.1007/s11336-021-09756-3>
- Hope, D., Kluth, D., Homer, M., Dewar, A., Goddard-Fuller, R., Jaap, A., & Cameron, H. (2025). Exploring the use of Rasch modelling in “common content” items for multi-site and multi-year assessment. *Advances in Health Sciences Education*, 30(2), 427–438. <https://doi.org/10.1007/s10459-024-10354-y>
- Hu, W., & Adey, P. (2002). A scientific creativity test for secondary school Student. *Research Report*, 24(4), 389–403. <https://doi.org/10.1080/09500690110098912>
- Irey, R., Blum, A. M., Carpenter, E., Miller, Z., Vonk, J. M. J., & Gorno-Tempini, M. L. (2026). Backward construct mapping in practice: using rasch measurement to reveal the internal structure of inflectional morphological processing in students with dyslexia. *Journal of Educational Measurement*, 63(2). <https://doi.org/10.1111/jedm.70047>

- Iriyadi, D., Asdar, A. K., Afriadi, B., Samad, M. A., & Syaputra, Y. D. (2024). Analysis of the Indonesian version of the Statistical Anxiety Scale (SAS) instrument with a psychometric approach. *International Journal of Evaluation and Research in Education (IJERE)*, 13(2), 672. <https://doi.org/10.11591/ijere.v13i2.26517>
- Irma, Z. U., Kusairi, S., & Yuliati, L. (2023). Strem Pbl with e-authentic assessment: Its Impact to students' scientific creativity on static fluid. *Jurnal Pendidikan IPA Indonesia*, 12(1), 80–95. <https://doi.org/10.15294/jpii.v12i1.40214>
- Jin, K.-Y., & Eckes, T. (2025). When raters generalize: Examining sources of halo effects with mixture Rasch facets models. *Behavior Research Methods*, 57(5), 149. <https://doi.org/10.3758/s13428-025-02667-6>
- Julianto, V., Sumintono, B., Almakhi, N. P. Z., Avetazain, H., Wilhelmina, T. M., & Wati, D. A. (2025). Academic Motivation Scale's psychometric attribute: analysis using Rasch measurement model. *Current Psychology*, 44(1), 114–124. <https://doi.org/10.1007/s12144-024-07142-7>
- Kangwanrattanakul, K., & Kulthanachairojana, N. (2024). Modern psychometric evaluation of Thai WHOQOL-BREF and its shorter versions in patients undergoing warfarin in Thailand: Rasch analysis. *Scientific Reports*, 14(1), 20639. <https://doi.org/10.1038/s41598-024-71048-4>
- Kiefer, C., & Sengewald, M.-A. (2025). Mining exceptional Rasch models. *Behaviormetrika*, 52(2), 361–391. <https://doi.org/10.1007/s41237-024-00251-4>
- Leung, K., Trevena, L., & Waters, D. (2018). Content validation of the evidence-based nursing practice assessment tool. *Nurse Researcher*, 26(1), 33–40. <https://doi.org/10.7748/nr.2018.e1544>
- Li, Y., Zhang, J., Sheng, Y., Yao, J., Zhang, X., & Xu, X. (2025). Rasch analysis and application research of the Chinese version of the Skidmore anxiety stigma scale: a cross-sectional study. *Scientific Reports*, 15(1), 41412. <https://doi.org/10.1038/s41598-025-25301-z>
- Liu, F., Qu, S., Fan, Y., Chen, F., & He, B. (2023). Scientific creativity and innovation ability and its determinants among medical postgraduate students in Fujian province of China: a cross sectional study. *BMC Medical Education*, 23(1), 1–13. <https://doi.org/10.1186/s12909-023-04408-9>
- Liu, Q., Wald, N., Daskon, C., & Harland, T. (2024). Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers. *Innovations in Education and Teaching International*, 61(4), 802–814. <https://doi.org/10.1080/14703297.2023.2222715>
- Lu, J., Wang, X., & Zhang, J. (2025). Joint modeling with generalized item response theory model family and response time model: Enhancing model structural flexibility and data-fitting adequacy. *Behavior Research Methods*, 57(12), 330. <https://doi.org/10.3758/s13428-025-02855-4>
- Naidoo, M. (2023). Twenty-First Century Learning (21 CL) - South African Private Secondary Schools in KwaZulu-Natal. *International Journal of Learning, Teaching and Educational Research*, 22(5), 26–57. <https://doi.org/10.26803/ijlter.22.5.2>
- Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of Rubrics on Academic Performance, Self-Regulated Learning, and self-Efficacy: a Meta-analytic Review. *Educational Psychology Review*, 35(4), 113. <https://doi.org/10.1007/s10648-023-09823-4>
- Pelouto, F., Papri, N., Haagsma, J. A., Cornblath, D. R., Nobile-Orazio, E., Wieggers, E. J. A., Harbo, T., Rajabally, Y. A., van Doorn, P. A., Gorson, K. C., Terwee, C. B., Jacobs, B. C., Horton, M. C., Antonini, G., Arends, S., Attarian, S., Barroso, F. A., Bateman, K., Benedetti, L., ... van Woerkom, M. (2026). Measurement properties of

- the Inflammatory Rasch-built Overall Disability Scale (I-RODS) in patients with Guillain–Barré syndrome. *Journal of Neurology*, 273(2), 164. <https://doi.org/10.1007/s00415-026-13684-6>
- Pinar, F. I. L., Panergayo, A. A. E., Sagcal, R. R., Acut, D. P., Roleda, L. S., & Prudente, M. S. (2025). Fostering scientific creativity in science education through scientific problem-solving approaches and STEM contexts: a meta-analysis. *Disciplinary and Interdisciplinary Science Education Research*, 7(1). <https://doi.org/10.1186/s43031-025-00137-9>
- Prahani, B. K., Suprpto, N., Rachmadiarti, F., Sholahuddin, A., Mahtari, S., Suyidno, & Siswanto, J. (2021). Online Scientific Creativity Learning (OSCL) in Science education to improve students' scientific creativity in covid-19 pandemic. *Journal of Turkish Science Education*, 18(Special Issue), 77–90. <https://doi.org/10.36681/tused.2021.73>
- Shaw, A., Kapnek, M., & Morelli, N. (2021). Measuring creative self-efficacy: An Item response theory analysis of the creative self-efficacy scale. *Frontiers in Psychology*, 12. <https://doi.org/10.3389/fpsyg.2021.678033>
- Sumo, M., & Soeharto, S. (2025). Fostering Scientific creativity through feasibility project-based learning in undergraduate physics. *J.Sci.Learn.2025*, 8(2), 191–205. <https://doi.org/10.17509/jsl.v8i2.83225>
- Touzani, R., Rouquette, A., Schultz, E., Allaire, C., Carrieri, P., Mancini, J., & Hardouin, J.-B. (2024). Psychometric validation of the French version of two scales measuring general (HLS19-Q12) and navigational (HLS19-NAV) health literacy using the Rasch model. *BMC Public Health*, 24(1), 3079. <https://doi.org/10.1186/s12889-024-20504-x>
- Tsai, Y. J., Chen, C. C., Hsu, Y. Y., & Chen, C. C. (2025). Validation of the uncertainty stress scale-high-risk pregnancy - Chinese brief version: Rasch analysis. *BMC Pregnancy and Childbirth*, 25(1), 6. <https://doi.org/10.1186/s12884-024-07078-7>
- Tutz, G. (2025). *The Binary Rasch Model* (pp. 9–35). https://doi.org/10.1007/978-3-031-87271-6_2
- Uba, A. R., & Khairani, A. Z. (2024). Assessing Item reliability, differential item functioning, and wright map analysis of the GSP122 Test at a Public University in Nigeria. *Journal of Education For Sustainable Innovation*, 2(2), 107–120. <https://doi.org/10.56916/jesi.v2i2.934>
- Xu, S., Reiss, M. J., & Lodge, W. (2024). Comprehensive Scientific Creativity Assessment (C-SCA): A New approach for measuring scientific creativity in secondary school students. *International Journal of Science and Mathematics Education*, 23(2), 293–319. <https://doi.org/10.1007/s10763-024-10469-z>
- Xu, S., Reiss, M. J., & Lodge, W. (2025). Comprehensive Scientific Creativity Assessment (C-SCA): A new approach for measuring scientific creativity in secondary school students. *International Journal of Science and Mathematics Education*, 23(2), 293–319. <https://doi.org/10.1007/s10763-024-10469-z>
- Yuanyuan, P., Alias, B. S., Mansor, A. N., & Hamid, M. R. A. (2025). Development and validation of the principals' digital leadership instrument using Rasch measurement model. *International Journal of Evaluation and Research in Education (IJERE)*, 14(3), 1577. <https://doi.org/10.11591/ijere.v14i3.32214>
- Zaky Adhitya Ginanjar Sasmita, Wahono Widodo, & Sifak Indana. (2021). Contextual based learning media development to train creative thinking skill in primary school. *IJORER: International Journal of Recent Educational Research*, 2(4), 468–476. <https://doi.org/10.46245/ijorer.v2i4.124>
- Zhao, Y., Luo, Y., Miao, Y., Li, J., Hou, Q., Dallakoti, N., & Wu, Y. (2025). Rasch analysis

of the mobile version of the eHealth literacy scale for cardiovascular disease patients. *Scientific Reports*, 15(1), 34093. <https://doi.org/10.1038/s41598-025-14198-3>