Journal of Science Education

ISSN 0124-5481

Development and Verification of Scientific Thinking Test Tool for Ninth-grade Students Based on Physical Situation

Xiaoyan Wu[a], Jing He[b*], Di Yin[c], Xiaofeng Hu[d], Zhuoyang Wang[e]

[a] *School of Physics and Electronic Technology, Liaoning Normal University, Liaoning 116020, P.R. China*
[b] *The Affiliated High School of Peking University, Beijing 100086, P.R. China*
[c] *Guanzhuang Branch of No.80 High School of Beijing, Beijing 100024, P.R. China*
[d] *Beijing Bayi School, Beijing 100080, P.R. China*
[e] *Beijing Shangdi Experimental School, Beijing 100085, P.R. China*

ARTICLEINFO

Keywords:
scientific thinking
Ninth grade students
Rasch model
quality analysis

ABSTRACT

Scientific thinking is an important part of the core literacy in the new curriculum standard of 2022. This study developed a tool to evaluate the scientific thinking of ninth-grade students. Based on the Rasch model, the overall quality of the tool was analyzed from the average difficulty estimation, error, data and model fitting index, separation, reliability and White diagram. The results show that the overall reliability and validity of the evaluation tool are high, the fitting degree and separation degree of the test questions meet the requirements, and the error is within the acceptable range. The White chart shows that the students' scientific thinking level is weak. Overall, the evaluation tool meets the requirements and has strong reliability. The test quality analysis based on Rasch model also provides a reference for the optimization of test tools.

1.The question is raised

The 2022 edition of junior high school physics curriculum standard points out that the core literacy of physics is mainly composed of four elements: physical concept, scientific thinking, scientific inquiry and scientific attitude and responsibility. Among them, scientific thinking mainly includes scientific reasoning, scientific argumentation, model construction, questioning and innovation, etc. It is a cognitive way to judge and explain the nature, laws and their relationships of objective things based on factual evidence and scientific concepts through scientific reasoning and argumentation. Scientific thinking runs through the process of scientific inquiry, which uses various thinking methods such as induction and deduction, comparison and classification, analysis and synthesis, abstraction and generalization and critical thinking. Scientific thinking ability is cultivated from childhood. In the study of physics subjects in middle schools, many laws and concepts are the products of scientific thinking, such as the law of conservation of mechanical energy, mechanical movement and so on. Scientific thinking is the ability that junior high school students must have to solve complex physical problems and events in their lives. Therefore, middle school students should pay attention to developing their own scientific thinking while learning physical knowledge. However, there are few studies on students' scientific thinking ability in previous studies. This study has developed test questions and scales suitable for measuring middle school students' scientific thinking ability. The research samples are mainly concentrated in the ninth grade students, whose reading and writing skills are well trained, and their logical thinking ability is in a period of rapid development. Therefore, we have developed a set of paper-and-pencil tests for evaluation, and tested the evaluation tools by Rasch, trying to measure the ninth grade students' scientific thinking ability and preliminarily verifying the rationality of the test questions.

2.The research design

(1) The research object

* Corresponding author:Jing He
Email: hejing0561@126.com
Accepted 1 October 2022, Available online 10 July 2023

174 valid questionnaires were recovered, with an effective rate of 97.21%. Of the official 174 subjects, 48% are boys and 52% are girls, with a ratio of male to female close to 1:1. The measured information of the samples is shown in Table 1 below. The middle school has excellent teaching quality and abundant students, and there are some differences among students. During the test, the subjects have learned all the knowledge involved in the test questions, which meets the knowledge base of the scientific thinking ability test under the background of physics.

Table 1 Measured information of samples

classes	schoolboy	girl student	overall number of people
1	17	22	39
2	17	20	37
3	13	21	34
4	23	18	41
5	2	2	4
6	12	7	19

(2) Research tools

In this study, the corresponding data were obtained through standardized physical examination and scientific thinking test papers. Standardized physics examination is used to collect students' physics scores, and scientific thinking test paper is used to collect students' scientific thinking level. The preparation process of the test paper is as follows: (1) The questions for examining scientific thinking are selected from the 2019-2022 middle school physics test papers collected from official website, the four major test centers in Britain, and official website, the education center in 50 States in the United States; (2) Classify and summarize the topics according to the knowledge points investigated, and screen out high-frequency test sites; (3) Comparing the high-frequency test sites with the Chinese compulsory education science curriculum standards, screening out the test questions with overlapping knowledge points, and after four rounds of discussion by the physical education expert group and the front-line teachers, selecting "electricity" as the research topic; (4) Three contextualized scientific thinking ability test questions were selected and adapted under the knowledge of electricity, and a scientific thinking test paper was formed to examine scientific reasoning, scientific argumentation and scientific modeling, and each dimension was evaluated from three sub-dimensions, with a total of nine evaluation items. Finally, through four rounds of discussion between the physical education expert group and the front-line teachers, the scientificity and standardization of the questionnaire are ensured.

After the formal test, data were collected, Rasch model was used as the measurement model, and the reliability, difficulty, unidimensionality and model fitting degree of the test questions were analyzed by Winsteps software to test the quality of the test questions. Using SPSS27.0 to analyze the collected data, we can understand the present situation of students' scientific thinking ability and provide reference for teaching strategies to cultivate students' scientific thinking ability.

3.Data analysis and results

(1) The overall quality analysis of testing tools

174 sample data were imported into Winsetps software for calculation, and there were no missing values (unanswered) in the sample data. All the answers of 174 Person were regarded as valid, and 9 evaluation item were estimated by the software. Rasch model mainly analyzes the overall quality of tools from several aspects, such as average difficulty estimate (Measure), Error (error), fitting index between data and model (Infit and Outfit), Seperation and Reliability. See figure for specific results.

	TOTAL SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	6.4	9.0	-1.04	.68	.99	.0	.95	.0
S.D.	4.1	.0	1.56	.19	.46	1.0	.60	1.0
MAX.	17.0	9.0	4.01	1.20	2.57	3.2	3.68	4.5
MIN.	1.0	9.0	-3.49	.53	.24	-2.9	.15	-2.5
REAL RMSE	.75	TRUE SD	1.36	SEPARATION	1.81	PERSON RELIABILITY	.77	
MODEL RMSE	.70	TRUE SD	1.39	SEPARATION	1.97	PERSON RELIABILITY	.80	
S.E. OF PERSON MEAN = .13								

Figure 1: Overall Quality Analysis

The analysis of the data results is as follows: the average difficulty estimate of the project will be automatically set to 0 by the Rasch model, and if the difficulty estimate of the subject is greater than 0, it means that the ability of the subject in the test is high or the difficulty of the project is low; If the difficulty value of the subject is less than 0, it means that the ability of the subject in the test is low or the difficulty of the item is high. As can be seen from the figure, the average ability of the subjects is -1.04logits, which shows that the students' ability to be tested is lower than the difficulty of the project, the test questions are difficult for the students, and the overall scientific thinking ability of the students is weak. From the reliability point of view, in the Rasch model, reliability is the accumulation of information functions of all items in the test, which is based on the test information function。 Generally speaking, the reliability is between 0~1, and 0.7~0.8 is the best. After testing, the reliability of this study is 0.77, the reliability index is good, and it has a high degree of credibility.

Infit and Outfit reflect the fitting degree between the actual observation and the theoretical model through two indicators: mean square (MNSQ) and standardization (ZSTD). Outfit MNSQ is the mean square of residuals, Infit MNSQ is the weighted mean square of residuals, and the ideal value of MNSQ is 1, which means that the actual data is completely fitted with Rasch model, and the acceptable range is 0.7~1.3. Infit and Outfit

The standardized forms of ZTSD are Infit ZSTD and Outfit ZSTD, both of which obey T distribution. The ideal value of ZTSD is 0, and the acceptable range is -2~+2. From the point of fitting degree, the average (infit)MNSQ=0.99 and (outfit)MNSQ=0.95, which are within the standard range (0.7-1.3) and close to 1. The closer to 1, the better the fitting degree, indicating that the test is close to the real level of students. The average value (infit)ZSTD=0.00, and the average value (outfit)ZSTD=0.00, which is within the standard range (-2.0, +2.0) and close to 0. On the whole, it shows that the samples are representative. The evaluation tool of scientific thinking ability designed in this study can effectively evaluate the scientific thinking ability of subjects.

(2) One-dimensional analysis

Unidimensionality is a basic assumption of Rasch model. Unidimensionality means that each evaluation item of the evaluation tool is measuring the same potential trait, and there is a main factor that can explain the variance in the sample response, which is the single main ability or single main potential trait measured by the test. One-dimensional does not mean that students only use one kind of ability to answer questions in the test, but it is mainly one kind of ability that affects students' response. For example, in the "scientific thinking ability" test designed in this study, students' "scientific thinking ability" plays a leading role.The standard residual diagram shows the degree of interpretation of all items, with the vertical axis showing the load of factors and the horizontal axis showing the difficulty of items. Except for items 1 and 3, the factor loads of the other seven items all fall between (-0.4, +0.4), which meets the requirements. On the whole, the test meets the assumption of one

dimension.

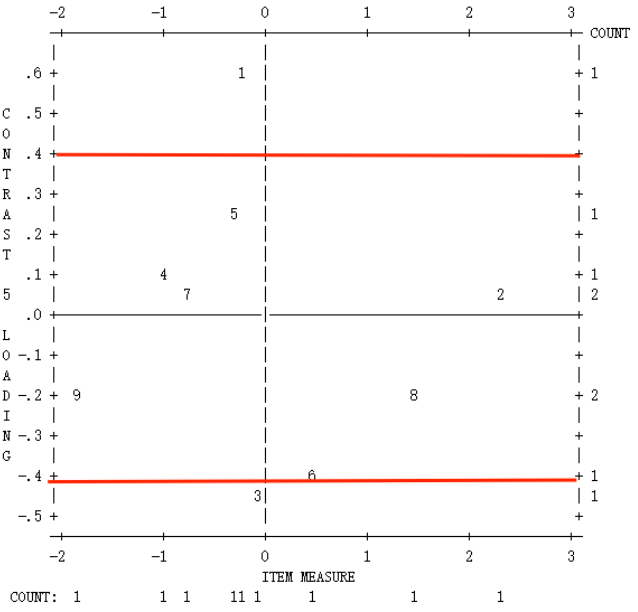


Figure 2: Standard residual diagram

(3) Subjects-project correspondence analysis

The subject-project diagram can measure the distribution of students' ability level, compare the subject's ability and the difficulty of the project on a ruler, and show the relationship between the subject and the project. In the figure, each "#" represents 6 subjects, and each "."represents 1-5 subjects. On the left side, the distribution of students is tested, and the ability of subjects is gradually weakened from top to bottom, while on the right side, the distribution of project difficulty is decreasing from top to bottom. The middle vertical line is logit scale, and the numbers 2, 1, 0, -1 and -2 represent logit scores. From the diagram, the sub-dimension "designing experiments and generating data" under the scientific reasoning dimension is the most difficult for the subjects, and the sub-dimension "modeling meta-cognition and meta-modeling knowledge" under the scientific modeling dimension is the simplest. The range of subjects' ability level is greater than the difficulty distribution of test questions, and the range of subjects' scientific thinking ability is wider, with significant differences, and the subjects' scientific thinking ability is at the lower-middle level.

Figure 3: Project-Subjects Correspondence (White Diagram)

(4) Project fitting analysis

Rasch project fitting statistics show the fitting degree between the measured data and the theoretical model (Linacre, 2002), which can further test the fitting situation of each project and be used to identify and screen the projects with poor fitting. Rasch model mainly reflects the quality of each evaluation project from the estimated project difficulty (Measure), model standard error (Model S.E), fitting index (Infit and fit) and point-measurement correlation (PTMEA CORR).

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM	G
2	36	174	2.29	.22	1.16	1.1	1.06	.3	.48	.52	81.7	80.7	SJSY	0
8	51	174	1.48	.19	1.30	2.1	1.18	.9	.50	.58	71.1	75.5	MXJY	0
6	75	174	.48	.16	.53	-4.6	.43	-3.9	.76	.64	83.8	68.1	TLFB	0
3	102	174	-.06	.16	.66	-3.4	.61	-3.6	.80	.70	77.5	65.5	JSSJ	0
1	98	174	-.21	.15	.72	-2.5	.62	-2.2	.74	.68	70.4	63.8	TCWT	0
5	103	174	-.32	.14	.84	-1.3	.67	-1.9	.73	.69	69.0	63.4	SSLL	0
7	127	174	-.78	.14	1.52	4.1	1.52	2.9	.62	.73	54.2	60.6	MXGJ	0
4	140	174	-1.04	.14	1.26	2.2	1.37	2.2	.68	.74	45.8	59.3	GD	0
9	177	174	-1.85	.15	1.06	.6	1.11	.7	.76	.78	57.0	63.0	JMYRZ	0
MEAN	101.0	174.0	.00	.16	1.01	-.2	.95	-.5			67.8	66.7		
S.D.	41.3	.0	1.20	.03	.31	2.8	.36	2.3			12.2	6.7		

Figure 4: Item Fitting Degree

The data in the table are fitting data of evaluation items, which are arranged according to the estimated difficulty value from high to

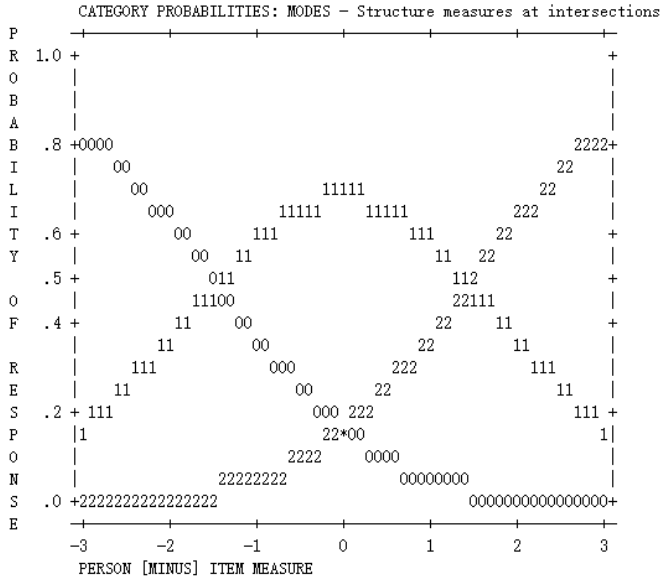
low. The difficulty values of nine items are in the range of -1.85~+2.29, and the item with item number 2 is the most difficult and the item with item number 9 is the least.

Infit and Outfit reflect the fitting degree between the project and Rasch theoretical model through two indicators: Mean Square Fitting Statistics (MNSQ) and Standardization (ZSTD). Among them, Infit MNSQ refers to the weighted mean square residual, which is sensitive to the case data with the difficulty of the topic equivalent to the individual ability level; Outfit MNSQ refers to unweighted mean square residual, which is more sensitive to extreme (abnormal data); ZTSD is a standardized form of MNSQ, and the standardized Z value is more sensitive in large sample research. Generally speaking, the ideal value range of MNSQ is between 0.75 and 1.33 (Wilson, 2004)X9. In project analysis, the Outfit MNSQ index should be checked first, and then the Infit MNSQ index and the ZTD index should be looked at. For the MNSQ index and the ZTD index, if the value of the project MNSQ index is between 0.75 and 1.33 (some documents also relax the standard to As can be seen from the above table, the fitting of all 9 items of the evaluation tool is good.

Point-measure correlation coefficient (PT-Measure CORR) indicates the correlation between the score and the total score of a certain item, and can reflect item polarity of the item, which is an index to measure the effect of item discrimination. The range of the point-measurement correlation coefficient is 0 ~ 1. The closer to the ideal value of 1, the greater the correlation, indicating that the project can well distinguish the ability of the subjects. If the value is less than 0, it indicates that there may be problems in the project. The numbers in this column in this table are all between 0 and 1. All the items in the test questions are consistent, indicating that the validity of the evaluation tool is good.

(5) Analysis of grading structure

The grading structure chart can reflect the structural characteristics of each item's grading, and check whether the pre-established grading standard is reasonable. All items in the evaluation tools used in this study are scored by 0, 1 and 2. Take the grading structure diagram of Item1' s three grades of scoring as an example for analysis, as shown in the figure.



The abscissa of the grade structure chart shows the difference between the scientific thinking ability of the subjects and the difficulty value of the project, and the ordinate shows the probability that the subjects get a certain score (0, 1, 2). Threshold is the intersection of curves in the figure, which corresponds to the same ordinate, indicating that the subjects with the level of scientific thinking ability corresponding to this point have the same scoring probability on these two scores. According to the hierarchical structure diagram of Item1, the score probability of the 0-point curve is lower and lower with the improvement of the scientific thinking ability of the subjects, and the score probability is 0 when the difference between the ability of the subjects and the difficulty of the project is greater than 1.5 or so; On the contrary, the score probability increases with the improvement of the scientific thinking ability of the subjects, and the score rate is 0 when the difference is less than -1.5. With the improvement of subjects' scientific thinking ability, the score probability of 1-point curve first rises and then falls, and there is a "peak" in the range of-0.5 ~+0.5 between subjects' ability and project difficulty, and the peak value is higher than that of 0 and 2 curves, and the three curves in the diagram cover a wide range of abilities. From the above analysis, we can know that the three scores of 0, 1 and 2 can represent a certain category well, which shows that the pre-established scoring standard is reasonable. Based on the above ideas, this study analyzes the grading structure of the remaining eight projects one by one, and finds that the grading standards set are close to the actual situation, and the formulation of the grading standards is scientific and reasonable.

Figure 5: Item1 rating scale structure

To sum up, all the indicators of the evaluation tool reflect good characteristics, which means that the designed evaluation tool for scientific thinking ability can be used to measure and evaluate the scientific thinking ability of the subjects.

4.Discussion and reflection

In this study, the scientific thinking ability test questions based on Rasch model are of high quality, and the difficulty of the test questions has a good fitting degree with students, which can be used to test students' scientific thinking ability, indicating that the test questions compiled this time can test students' potential characteristics. In order to make the test paper have high-quality reliability and validity, not only high-quality test questions, but also scientific grading rules and reasonable teaching cooperation are very important. When using the Rasch model to analyze the test paper quality, we should first judge whether the test meets the applicable conditions of the Rasch model, and conduct multiple tests when evaluating the sample quality, and also consider whether the tested content is fair and universal to the test group. The difficulty level of the test questions should also match the level of the test samples, so as to ensure the effectiveness of the test tools.

References

Luo Zhaosheng. Theoretical basis of project response [M]. Beijing: Beijing Normal University Press, 2012.

BOND T G, FOX C M. Applying the Rasch model: Fundamental measurement in the human sciences[M]. 3rd ed. NY: Routledge,2015.