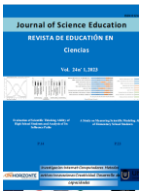
Journal of Science Education

ISSN 0124-5481



Development of a Scientific Thinking Assessment Tool for Sixth Graders Based on Chinese 2022 Edition of the New Academic Standards

Wenyu Yuan[a], Lu Lu[a], Ziwei Liu[b], Wenke Liu[a], Xuewei Tian[a*]

[a] *Normal College & School of Teacher Education, Qingdao University, Qingdao 266071, P.R. China*

[b] *Faculty of Education, Beijing Normal University, Beijing 100875, P.R. China*

ARTICLEINFO

Keywords:
scientific thinking
scientific thinking assessment
tool development

ABSTRACT

The revision of Chinese Science Curriculum Standards for Compulsory Primary Schools (2022 Edition) is led by the core literacy that reflects the comprehensive nurturing value of the science curriculum. Scientific thinking is the core of the core literacy. Understanding scientific thinking is the key to in-depth understanding and effective implementation of the new standards. Assessing students' scientific thinking is conducive to understanding the level of scientific thinking in China. In this study, the scientific thinking assessment was carried out on Grade 6 students in Jinzhou City, Liaoning Province, by using the developed scientific thinking assessment questions for elementary school, and the quality of the test papers was analyzed with the help of Winstepts3.81.0 software. The results showed that the overall quality of the assessment test questions and the quality of each test question were relatively good, close to the assessment objectives, and the fit of the test questions was good, which could accurately assess the scientific thinking level of the students.

1.Introduction

Chinese Science Curriculum Standards for Compulsory Primary Schools (2022 Edition) states that the science curriculum should cultivate students' core literacy, which mainly refers to the correct values, necessary character and key abilities that students gradually form in the process of learning the science curriculum to adapt to their lifelong personal development and the development of the society, and includes scientific concepts, scientific thinking, inquiry and practice, and attitudes and responsibilities. Scientific thinking is a way of recognizing the essential attributes, internal laws and interrelationships of objective things from a scientific perspective. Scientific thinking is the core of core literacy, and the formation of all core literacy depends on active thinking. The new standard integrates various forms and methods of scientific thinking, highlights typical types of scientific thinking, and proposes that scientific thinking includes model construction, reasoning and argumentation, and innovative thinking. In the past thirty years, foreign psychologists' research on the development of scientific thinking has focused on the three skills of formulating hypotheses, designing experiments, and analyzing and interpreting evidence (Fan, 2022). It can be seen that scientific thinking covers the three thinking skills of reasoning, argumentation and modeling, which can also be regarded as the dimension that reflects the level of scientific thinking. Scientific thinking level, as an implicit trait, needs to be diagnosed and assessed through outward behaviors, i.e., it can be based on descriptions and analyses of students' responses when solving scientific problems, which can then be used to make reasonable inferences about students' scientific thinking level. In this study, we used elementary school science test questions to develop a test instrument to assess the level of scientific thinking. Assessment is measurement and evaluation, i.e., obtaining persuasive data and facts about students' academic performance based on tests and making judgments about students' academic performance accordingly. The ability to use scientific and accurate assessment tools will have a direct impact on the evaluation of students' scientific thinking level, so it is particularly important to analyze the quality of test paper

2.Assessment Tools and Testing Models

2.1 Scientific Thinking Assessment Tool Development Process

Scientific reasoning, argumentation and modeling are domain-attributed skills of scientific thinking and the main "doers" of scientific thinking practices in science. These three functions have been described as means of logical reasoning, tools of pedagogical reform and components of practical activities. In the process of scientific learning, they are not independent of each other, but rather a cyclical whole (Zheng, Zhou& Wang, 2023). The practice of teaching

* Corresponding author: Xuewei Tian
Email: tianxuemao@163.com
Accepted 1 September 2023, Available online 15 October 2023

scientific thinking stems from the exploration of scientific reasoning. Therefore, the evaluation tool developed in this study is based on the concept of comprehensively mobilizing students' scientific reasoning, scientific argumentation, and scientific modeling abilities, integrating the evaluation of the three abilities into the same test question. And the scale was set according to the three-level division criteria for the three sub-competencies and nine sub-dimensions of each test question.

The development of the assessment tool was carried out in strict accordance with scientific norms. Firstly, the dimensions of scientific thinking assessment were determined, and the principles and concepts of the development of the test paper were determined based on the target requirements for scientific thinking of sixth grade elementary school students in the Compulsory Education Primary School Science Curriculum Standards (2022). Secondly, we referred to the logic of writing test questions for elementary school science in Edexcel Examination Center of the United Kingdom and the states of New York and Texas in the United States, closely followed the content of the textbook of Science for Science Education Edition for Primary Schools, determined the observation points according to the specific learning tasks, constructed the assessment tools preliminarily, and organized in-house experts to review the test questions and to modify and improve the content and structure of the test questions according to the feedback. Pre-experimentation was again conducted on a small scale, and the structure and content of the tool were adjusted according to the pre-experimentation data. Next, a large-scale test was conducted in Jinzhou City, Liaoning Province, and the quality of the test tool was analyzed based on actual test data. It was finally determined that the assessment tool had a total of three major questions, including nine minor questions, all of which were expository in nature. The process of developing elementary school science test questions includes four processes: test paper retrieval, test question screening, test question changing and task question development.

The test paper retrieval process visited the official websites of each of the four major examination boards in the UK: the AQA (The Assessment and Qualifications Alliance), OCR (Oxford Cambridge and RSA Examinations), CIE (Cambridge International Examinations) and Edexcel's official websites as well as the official education websites of all 50 states of the United States of America, collected the science final exams and stage exam test papers for grades 4-6 for the years 2019-2022. The selection process of the test questions categorized the collected test papers according to the knowledge points examined in the questions, compared the content of the higher frequency of the exam with the content of the domestic six-year science curriculum standards for elementary school, screened the test questions for overlapping knowledge points, and selected the questionnaire questions to examine the topic of "Earth's Motion" after four rounds of discussions among the research group, expert group and front-line teachers. After four rounds of discussion among the subject group, the expert group and front-line teachers, the theme of the questionnaire was selected as "Earth's motion". Among the test questions on the "motion of the Earth", two were selected as the most capable of examining students' higher-order thinking ability with context creation, namely, Question 21 of the 2018 final test paper and Question 26 of the 2022 final test paper of fifth-grade science in Texas, U.S.A.; and the first question of the science test question in the questionnaire was adapted from Question 21 of the 2018 test paper mentioned above. The first question in the science test question part of the questionnaire was adapted from question 21 of the abovementioned 2018 test paper, and the second and third questions were adapted from question 26 of the abovementioned 2022 test paper; the three test questions in the task questionnaire were all discussed in four rounds by the subject team, the expert group, and the frontline teachers in order to ensure that the questionnaires were scientifically sound and standardized. The task questions were divided into three parts: the first part was the basic information about the students, the second part was the three test questions, and the third part was the investigation of various background information about the students. The questionnaire was answered in such a way that the next part of the questionnaire was displayed only after the previous part was completed.

2.2 Application of the Rasch model

Analyzing the quality of assessment tools is an important part of educational assessment. Once there are problems with an assessment tool, the results obtained based on the tool are bound to be inaccurate, incomplete or even questionable, which will inevitably lead to unfair evaluation. In order to realize the fairness of educational assessment, scientific analysis methods are indispensable. At present, the classical testing theory (CTT) is mostly used to analyze educational assessment in China, but the theory has many limitations, such as sample dependence, inaccuracy of reliability measurement, and neglect of test question response group type (Luo, 2012), which leads to the analysis of assessment based on CTT can not reflect the real level of the students better. In addition, CTT-based assessment is relatively powerless in dealing with the new college entrance examination "two exams in one year" measure, there is bound to be a difference in difficulty between the two exams, and the fairness of the assessment results is highly controversial if the true scores are directly used to assign scores and compare them, which also leads to the emergence of some speculative behaviors, so that the results of assessment are no longer purely a reflection of students' abilities. purely a reflection of the student's ability.

Item response theory (IRT) can effectively make up for the shortcomings of CTT. There are ten kinds of statistical models constructed based on IRT, and Rasch model is one of the more widely used models. Rasch model is equivalent to one-parameter logistic model in terms of mathematical expression, and the debate about the relationship between the two persists (ANDRICH D, 2004), and the vast majority of researchers regard the two as the same model (Qi, 2002), and some researchers believe that the two are different in the sense of use. researchers believe that the two are different in the sense that they are used; Logistic model is used to describe and fit the data, and when the data fit is poor, the model will be modified or another model will be chosen directly; Rasch model, on the other hand, fits the model with the data, and when the data fit is poor, the poorly fitted items will be modified or deleted, and the model will be measured again in order to obtain a good model fit. It can be seen that the Rasch model, unlike other statistical models that attempt to fit existing data, inverts the relationship between the data and the model, requiring the assessment data to fit the ideal model, providing a more objective criterion for the modification of the assessment instrument, and resulting in an assessment that is more indicative of the relationship between the item's traits and the individual's underlying traits (COER, 2008).

The Rasch model is based on Item response theory, which states that the probability that a particular individual will respond in a particular way to a particular topic depends on a simple function between the individual's ability and the difficulty of that topic (Li, 2016).An important feature of the Rasch model is that the individual and the topic share a common ruler (logit), and the individual's ability and the topic difficulty will be converted to the same unit of measurement, which will enable a directly compare the relationship between individual ability and topic difficulty (Wu, Tian, Wang & Fan 2021). The likelihood of a respondent answering a question correctly depends on the comparison between their ability level and the difficulty of the question. The level of an individual's ability and the difficulty of the question directly determine whether the individual can answer the question correctly. When an individual's ability is higher than the difficulty of the question, the individual is able to answer the question correctly; conversely, he or she is not able to answer the question correctly (Bai, Zhu & Chen, 2019). The level of the respondent or the difficulty of the questions is reflected in the testing process, and only the part of the test questions that meets the requirements of the

various parameter indicators of the model can more accurately reflect the actual ability of the respondent, and these parameter indicators, including error, difficulty, reliability, fit, dimensionality and so on, is an important indicator that reflects the quality of the test questions.

With the continuous development and improvement of the application software for validating the Rasch model, scholars at home and abroad have been deepening their research on the application of the model in the field of education. Many educational researchers have based on the Rasch model to assess the teaching process and academic ability and achieved convincing results. Maja Planinic, a famous researcher in physics teaching, utilized the Rasch model to evaluate the conceptual test of mechanics (Maja Planinic, Lana Ivanjek & Ana Susac, 2010).Talib et al. (2018) took 150 sophomore students as research subjects and analyzed the learners' final exam response data in a course through Rasch model, and proposed a procedural method that can effectively measure the reliability and validity of the test instrument. Based on the analysis of PISA (Program for International Student Assessment) test questions using the Rasch model, Wang Lei (2007) discusses the reference value of Rasch, an objective equidistant measurement scale, for the improvement of China's educational evaluation and psychometrics. Peng et al. (2022) carried out a quality analysis of ICT primary test papers using the Rasch model and found that the test results had high reliability, the difficulty of the questions in the test papers matched the ability level of the test takers, and they could effectively measure and differentiate the intercultural competence of the test takers. Zheng et al. (2019) conducted the level construction, the development and optimization of the test instrument, and the analysis of the assessment results of "Physical Science Argumentation Ability". The study shows that the revised test instrument meets the relevant quality indicators of the Rasch model and has credibility.

3. Assessment methodology and quality analysis

3.1 Sample composition and recall

Eight districts in Jinzhou City, Liaoning Province, namely Beizhen District, Gaoxin District, Guta District, Heishan District, Linghai District, Linghe District, Taihe District, and Yixian District, were selected for this assessment, which was distributed using online questionnaires and voluntarily answered by the city's sixth-grade students. There are currently a total of 16,729 sixth-grade students in Jinzhou City, and the questionnaires returned for this assessment are shown in Table 1. The total number of online questionnaire fillers was 9737 students, and the total number of valid questionnaires recovered was 8693, with a questionnaire fill-in rate of 58.2% and an effective recovery rate of 89.28%. This study randomly selected one-half of the overall sample from the total sample size in each region except Marina City and the city as the research object, totaling 4,415 copies.

Table 1 Online Questionnaire Returns for Scientific Thinking Task Questions for Sixth Graders

District	Number of students	Number of recoveries (copies)	Effective questionnaires (copies)	Questionnaire completion rate	Questionnaire representation rate	Effective recovery rate
Linghe	2246	1223	1061	54.45%	47.24%	86.75%
Guta	1301	853	816	65.56%	62.72%	95.66%
Taihe	698	498	472	71.34%	68.51%	94.78%
Gaoxin	649	649	452	72.57%	72.57%	69.65%
Linghai	2449	1901	1686	77.62%	68.84%	88.69%
Beizhen	2757	1576	1413	57.16%	51.25%	89.66%
Heisha n	270	1929	1897	71.36%	70.18%	98.34%
Yixian	2055	1138	1030	55.38%	50.12%	90.51%
add up the total	16729	9737	8693	58.20%	51.96%	89.28%

3.2 Quality testing

In this study, SPSS 26. 0 and Winsetps3.81.0 software were used to analyze the test data by Rasch modeling. Each question measured students' scientific thinking in three dimensions: scientific reasoning, scientific argumentation, and scientific modeling, and the codes for the indicators of each dimension are shown in Table 2. When using the Rasch model to analyze and process the data, an a priori condition needs to be satisfied - the input data must be fitted to the Rasch model. However, in practice it is difficult to achieve a perfect fit with the model, taking this into account, the Rasch model only requires the data to be within the fit range.

Firstly, the overall quality of the test questions was tested, 4415 data were imported into Winsetps3.81.0 for the calculation, and it was found that there were no missing values (unanswered), 47 subjects' (Person's) responses were regarded as very low scoring, and the remaining 4368 subjects' (Person's) responses were regarded as valid, and 27 items (Item) were estimated by the software. The Rasch model is mainly from the The overall quality of the test questions was analyzed in terms of Mean Difficulty Estimate (Measure), Error, Data and Model Fit Index (Infit and Outfit), Seperation, and Reliability. In the Rasch model, the average difficulty estimate of the question (Item) is set to 0, so the estimate of the subject's (Person's) Measure is actually the average ability value of the student. The overall profile of the students and the overall profile of the Measure test questions are shown in Table 3. The average ability of the students in this study was -0.58, which is lower than the item difficulty value, indicating that the test questions as a whole were difficult for the students, but the difference was not huge. This indicates that the assessment test questions fit the students' literacy level well and were appropriate for this sample. Error represents the difference between the theoretical model and the actual observations, and the subject error (0.36) and item error (0.03) are both relatively close to 0, indicating that the data have a high degree of reliability and that the observations obtained through the preliminary assessment instrument can reflect the students' scientific competence in a more realistic way. Infit and Outfit represent the fit between the theoretical model and the actual observations, including the MNSQ and the MNSQ. The data in Table 3 show that the MNSQ and ZSTD of students and questions are very satisfactory, indicating that the observed values of the assessment tool fit well with Rasch's theoretical model and are close to the real level of students. The indicators of model reliability in Rasch analysis mainly include Separation index and internal reliability coefficient. When the Separation Index is greater than or equal to 2, the test can better distinguish between different levels of ability or difficulty; when the internal reliability coefficient is greater than or equal to 0.8, the measurement results have a high degree of reliability (Bond & Fox 2013). The test taker separation index is 3.31 and the internal reliability coefficient is 0.92, which are both within the reference range, indicating that the ability levels of the test takers participating in the test have a high degree of differentiation (Engelhard,2013). The test question separation index was 13.41 and the internal reliability coefficient was 0.99, both within the reference range, suggesting that there is a multilevel differentiation of difficulty between test questions (Bond & Fox 2013). These results indicate that the scientific thinking task questions test results are highly reliable, measure the level of scientific thinking of test takers in a more comprehensive way, and can effectively differentiate between the abilities of test takers.

Table 2 Indicator coding for dimensions of scientific thinking task questions assessment

Scientific reasoning	Asking questions and making assumptions	first question	AA
		second question	BA
		third question	CA
	Designing Experiments and generating Data	first question	AB
		second question	BB
		third question	CB
	Interpreting data and drawing conclusions	first question	AC
		second question	BC
		third question	CC
Scientific Argumentation	Viewpoint	first question	AD
		second question	BD
		third question	CD
	Factual and theoretical basis	first question	AE
		second question	BE
		third question	CE
	Reasoning and rebuttal	first question	AF
		second question	BF
		third question	CF
Scientific modeling	Model construction and use	first question	AG
		second question	BG
		third question	CG
	Model testing and correction	first question	AH
		second question	BH
		third question	CH
	Modeling metacognition and metamodeling knowledge	first question	AI
		second question	BI
		third question	CI

Table 3 Analysis of overall quality of scientific thinking task questions

	Measure	Error	Infit		Outfit		Separation	Reliability
			MNSQ	ZSTD	MNSQ	ZSTD		
Person	-.58	.36	.99	.0	1.05	.0	3.31	.92
Item	.00	.03	1.00	-1.8	1.05	-.2.2	13.41	.99

Second, dimensionality is one of the basic assumptions of the Rasch model, i.e., that subjects' performance on a particular item can be attributed to a single variable (knowledge, ability, personality trait, etc.) and that the effects of other factors on subjects' performance can be ignored (Jan-Eric Gustafsson, 1980). Therefore, the dimensionality test is a necessary step in the measurement analysis using the Rasch model. In the Rasch test, the standardized residual plot, which determines whether there are other factors influencing the subjects' responses, is used to conduct the dimensionality test. The horizontal coordinate of the residual plot indicates the item difficulty, and the vertical coordinate is the value of the correlation between the item scores and the possible influencing factors. As shown in Figure 1, the upper and lower case letters in the plot represent the 27 evaluation topics, the horizontal coordinate represents the topic difficulty, and the vertical coordinate shows the topic loading coefficient, which is ideally between -0.4 and +0.4, and beyond which, it is considered that it does not satisfy dimensionality requirements. As can be seen from Figure 1, f (topic BG) and c (topic BB) in this test slightly deviate from the ideal range. The loading coefficients of the vast majority of topics fall between -0.4 and +0.4, which can be considered to fulfill the requirement of dimensionality.

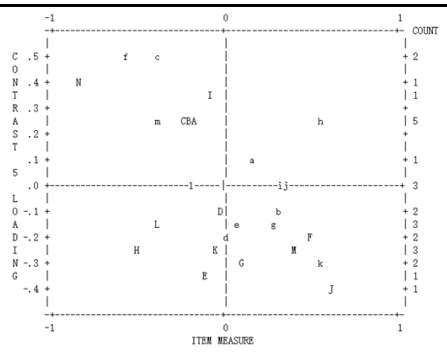


Figure 1 Dimensionality of test questions

Third, the test questions were analyzed for the structure of the topic rating scale, and the results are shown in Figure 2. The horizontal coordinate (PERSON[MINUS] ITEM MEASURE) in the graph represents the difference between the students' ability and the difficulty value of the questions, and the vertical coordinate (PROBABILITY OF RESPONSE) represents the probability of the students scoring 0, 1, and 2 points. At the Threshold position, where the curves cross in the graph, the same vertical coordinate corresponds, i.e., the student has the same probability of obtaining both scores. As can be seen from the result plots for the 27 dimensions, basically the rating scale category curves for each dimension have distinct peaks and are flat and cover a certain range in the horizontal coordinates, which is a good performance.

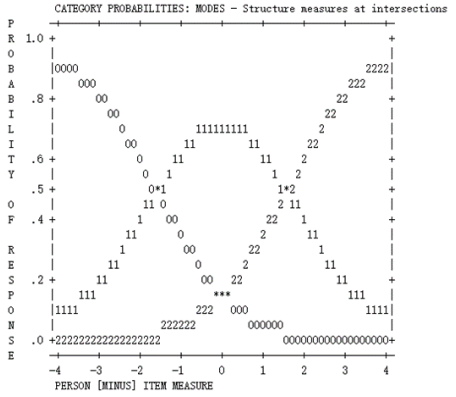


Figure 2 Structure of the grading scale for test questions

Fourth, test item fit testing. Table 4 shows the item fit obtained by importing the data collected in this assessment into winsteps software for parameter estimation, including item difficulty (Measure), standard error (S.E.), residual mean squares of Infit and Outfit, and correlation coefficients (CORR), of which Infit MNSQ is the weighted mean square of the residuals, which is more sensitive to the data that the item difficulty is comparable to the corresponding subjects' ability level, and Outfit MNSQ is the mean square of the standardized residuals, which is more sensitive to anomalous data, both of which are obtained by calculating the residuals. Typically, when the MNSQ value is between 0.50 and 1.50, the data fit the model to an acceptable degree as expected, and when the value is 1, the data fit the model perfectly.

The data in the table are the topic fit data for the test, organized in order of highest to lowest difficulty estimates. For the fit indices, OUTFIT is more important than INFIT and the OUTFIT index should be looked at first. Among these two indices, if the ZSTD exceeds the range of -2.0 to +2.0, it indicates a poor fit, but if the MNSQ is still between 0.7 and 1.3, such a fit is still acceptable, for example, in Table 4, CI, BH, BE, CB, CE, BI, BF, AA, BC, BD, AF, AB, CH, CF, BB, CG, CA, BG , CD, but AD, AH, AI, and AC were less well fitted. There were a large number of test questions, so the overall view was that the actual data from the

test questions fitted the model well. The standard error (S. E.) reflects the stability of the items in measuring the ability level of the subjects, and the smaller its value, the more stable the results of the items in estimating the ability level of the subjects. As can be seen from Table 4, the error values of all test items are within 0. 03, indicating that the test items are more stable in estimating the subjects' ability and the test instrument has high reliability. The correlation coefficient (CORR) indicates the proximity of the items to the measurement target, the higher the correlation coefficient, the closer the items are to the measurement target. As shown in Table 4, the correlation coefficients of all the items in this test are within a reasonable range, indicating that all the items can effectively measure the target.

Table 4 Fit of Test Questions

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
27	2665	4415	.61	.03	.90	-5.2	.83	-6.0	.63	.57	70.3	63.3	CI	0
17	3037	4414	.56	.03	1.18	8.2	1.20	8.4	.47	.56	61.9	67.1	BH	0
14	2790	4413	.55	.03	.93	-3.4	.88	-4.8	.62	.58	66.7	63.1	BE	0
4	2523	4415	.50	.02	2.06	9.9	4.29	9.9	.21	.59	46.1	63.7	AD	0
5	3315	4415	.40	.03	.99	-.4	.98	-1.0	.57	.56	69.8	68.8	AE	0
20	2965	4415	.35	.03	.83	-8.9	.75	-9.4	.67	.60	67.9	60.2	CB	0
23	2982	4415	.34	.03	.83	-8.7	.73	-9.9	.67	.60	67.6	60.2	CE	0
10	2931	4414	.31	.02	1.12	5.7	1.03	1.0	.58	.61	57.4	60.6	BA	0
18	3228	4414	.26	.03	.99	-.8	.95	-2.2	.61	.59	65.9	61.8	BI	0
15	3373	4414	.14	.03	.74	-9.9	.68	-9.9	.72	.60	69.2	60.6	BF	0
1	3483	4415	.10	.03	1.06	2.9	1.06	2.7	.57	.60	62.1	61.8	AA	0
12	3555	4414	.05	.03	.90	-5.6	.88	-6.0	.65	.59	67.1	62.2	BC	0
13	3553	4414	.00	.03	1.26	9.9	1.29	9.9	.50	.61	48.8	59.5	BD	0
6	3715	4415	-.03	.03	.75	-9.9	.73	-9.9	.71	.58	70.4	64.2	AF	0
2	3747	4415	-.06	.03	.96	-1.9	.95	-2.5	.61	.59	65.6	64.0	AB	0
26	3705	4414	-.08	.03	.88	-6.7	.90	-5.0	.65	.60	60.1	60.8	CH	0
21	3585	4415	-.12	.02	1.05	1.9	1.00	.0	.66	.66	61.3	63.0	CC	0
24	3727	4415	-.18	.02	.52	-9.9	.42	-9.9	.80	.65	74.8	60.0	CF	0
8	3846	4415	-.20	.03	1.45	9.9	1.48	9.9	.43	.62	49.2	58.8	AH	0
9	3857	4415	-.20	.03	1.36	9.9	1.38	9.9	.46	.62	51.2	59.4	AI	0
3	3861	4415	-.21	.03	1.44	9.9	1.47	9.9	.44	.62	49.1	58.5	AC	0
11	4107	4414	-.38	.02	.83	-9.4	.80	-9.0	.71	.64	64.8	57.8	BB	0
25	4133	4415	-.40	.02	.64	-9.9	.57	-9.9	.78	.65	68.4	58.4	CG	0
19	4136	4415	-.40	.02	.71	-9.9	.59	-9.9	.76	.67	68.3	60.2	CA	0
7	4287	4415	-.50	.03	.98	-1.0	.98	-.9	.58	.58	65.7	66.2	AG	0
16	4420	4414	-.57	.02	.79	-9.9	.74	-9.9	.74	.66	67.5	58.0	BG	0
22	4923	4415	-.84	.02	.89	-5.2	.84	-5.0	.71	.68	63.9	61.3	CD	0
MEAN	3572.2	4414.6	.00	.03	1.00	-1.8	1.05	-2.2			63.0	61.6		
S.D.	560.7	.6	.37	.00	.31	7.4	.69	7.3			7.6	2.7		

Finally, the white plot provides information about the distribution of item difficulty matched to students' ability levels, listing the locations of 4,415 students and 27 items on a common scale. The White plot is able to visualize the relationship between item difficulty and subject ability, subject to subject, and item to item on the same scale. As shown in Figure 3, the dotted line in the center is the logit scale, the numbers to the left of the dotted line represent difficulty or ability levels, the symbols "#" and "." represent the number of personnel, a "#" represents 20 candidates, a "." represents 1-19 students, while the right side of the dotted line shows the number of the test questions M is an abbreviation for Mean, which stands for average; S is an abbreviation for One Standard Error, which means one standard deviation from the mean; and T is an abbreviation for Two Standard Error, which means two standard deviations from the mean. Vertically, the left and right sides of the demarcation line show the distribution of students' ability level and the difficulty of the test questions from low to high, respectively, from bottom to top. Horizontally, the correspondence between the left and right sides of the demarcation line reflects the degree of matching between the students' ability level and the difficulty of the test questions. The intervals between subjects represent the differences in ability between each other, and the intervals between items represent the differences in difficulty between each other; the closer the distance, the smaller the differences. The content adequacy and validity of the test questions can be assessed from the distribution and ordering of the questions in the White's chart. If the distribution of the examinee-test question relationship in which examinees of different ability levels correspond one to one with the test questions of different difficulties, it means that the difficulty of the questions matches with the examinee's ability level, and the measurement is effective.

As can be seen from Figure 3, the distribution range of the difficulty of the questions in this test is about 2 logits, with a positive skewed distribution, and the average difficulty of the questions is at 0 logits, which indicates that the difficulty of the questions is moderate, and the questions BE, BH, and CI are the most difficult for, and the question CD is the easiest for, and the distribution range of the ability of the examinees is about 3.2 logits, and the mean value of the ability of the examinees is about -0.56 logit. overall, the mean value of the ability of the students is about -0.56 logit. In general, the gap between the mean value of students' ability and the mean value of the difficulty of the questions is large, indicating that there are more students with intermediate ability, the difficulty of the questions is concentrated in the middle level, and the distribution of the students' ability is more uniform, and the students' ability and the difficulty of the questions are basically matched successfully. The number of people in the interval where the subjects and the questions are not matched is larger, mainly in the difficulty -1, -2 and some students near difficulty 1 did not get the test questions matching their abilities. This suggests that there were a small number of candidates who were too low or too high in ability on this test, and that future tests will need to further optimize the candidate mix. There were no large noticeable gaps between questions, suggesting that the test questions were well distributed in terms of difficulty and had good content adequacy and validity. Overall, the overall match between the candidates' ability and the difficulty of the questions in this study was good, and the test was able to differentiate the candidates' level of intercultural competence, and the test was effective.

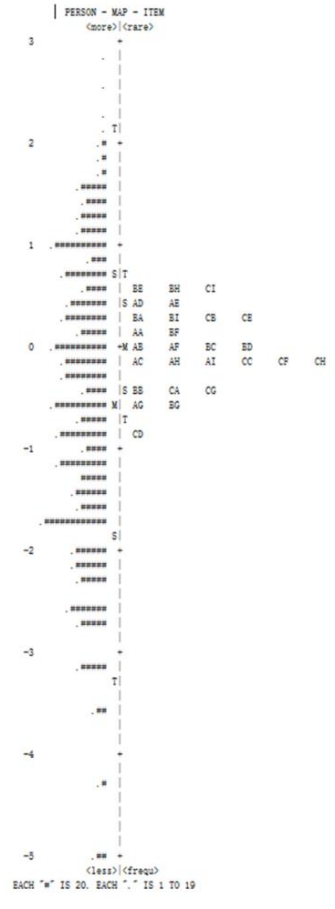


Figure 3 White plot

4. Quality analysis results and insights

4.1Results of the qualitative analysis of the assessment tool

This study analyzed the data of science thinking test questions from a total of 4,415 sixth-grade elementary school students in various districts and schools in Jinzhou City, Liaoning Province, through the Rasch model, and concluded that this elementary school science thinking test questions designed on the basis of the 2022 edition of the new standard had good reliability and differentiation (reliability of the questions was 0.99 and differentiation was 13.41, and the reliability of the subjects was 0.92 and differentiation was 3.31)). After analyzing the quality of the developed Science Thinking Assessment Tool for 6th grade elementary school through Winsteps3.81.0 software, the overall quality of the test questions, dimensionality, question fit, rating scale structure, and White's charts were analyzed in five areas, and it was found that the overall quality of the measurement tool was high, but some of the questions needed to be made adjustments at a later stage. The separation between questions in this study was large, the overall reliability was high, and the test paper could effectively measure subjects with different abilities. From the fit of the questions, it can be seen that T1-1 and T1-3 do not fit well with the Rasch model, indicating that there are some interfering factors in these questions in measuring the subjects' abilities corresponding to them, and modification of the formulation of the related questions should be considered at a later stage. Other than that, the rest of the questions fit well with the Rasch model. The overall quality analysis reflects that this test paper is difficult for the subjects, and at the same time there is a lack of questions matching the ability of some subjects, which should be considered to appropriately add some questions with different difficulties at a later stage. At the same time, the test questions generally showed good dimensionality, which can better reflect students' scientific thinking ability. To summarize, the overall quality of the assessment tool is high, the difficulty is high, and there is a certain degree of differentiation, the measured data of each topic is more in line with the data expected by the Rasch model, and the topics can all measure the target ability.

4.2inspiration

Cultivating students' core literacy requires a long-term and systematic educational process, a process from quantitative to qualitative change. Cultivating students' scientific thinking based on core literacy needs to be implemented in every lesson. The scientific thinking assessment tool developed based on the Rasch model can be used to detect the current status of students' scientific thinking level, and can objectively and independently evaluate students' thinking ability, providing a basis for teachers to carry out teaching based on core literacy.The Rasch model provides an intuitive and effective way to assess students' ability, and the use of the Rasch model can ensure the scientificity and validity of the assessment, and enhance the value of the assessment. The use of the Rasch model can ensure the scientific and effective assessment and enhance the value of the assessment. In order to ensure the accuracy of the analysis results, when compiling the test questions, we should ensure that each item meets the prior conditions of Rasch as much as possible, and develop the questions according to the test objectives, so as to build a teaching assessment tool suitable for China's local community. (Gao & Bai, 2021).

In order to formulate assessment questions based on scientific thinking, it is necessary, first of all, to shift the focus from memorization of fragmented knowledge to understanding and application of core scientific concepts. Secondly, it is necessary to create task situations that are closely related to students' lives, and in the process of solving problems, students' scientific thinking level should be comprehensively assessed. The assessment is based on test questions, which should be contextualized and diversified. Referring to the sample questions on the official websites of the four major examination bureaus in the UK, Edexcel, and the official websites of the 50 states in the U.S., the researchers should integrate the content of the test into rich contextualized materials and combine them with pictures, tables, or animations, etc., so as to make it easier for students to comprehend the questions and at the same time to increase the interest of the test questions.

"Assessment for teaching, assessment for learning", the test question assessment based on the Rasch model can provide a practical method for teachers to evaluate students more reasonably in the future, so as to take targeted measures to improve the scientific thinking level of primary school students, which is conducive to the in-depth implementation of the curriculum teaching reform. Depending on students' deficiencies at the level of scientific thinking, teachers can choose corresponding teaching methods and strategies. Higher-order thinking needs to be based on the achievement of lower-order thinking, so teachers need to gradually raise the cognitive level of learners in order to realize meaningful learning.

References

Science Curriculum Standards for Compulsory Education [M]. Ministry of Education of the People's Republic of China. Beijing Normal University Press.2022.

Wang J Y, Zhou D H, Yang Y, Ke L & Tian X W.(2023). Scientific higher order thinking: Connotation value, structure function and practice Approach. Modern distance education (02), 11 to 18, doi: 10.13927 / j.carol carroll nki. Yuan. 20230522.002.

Fan Z. Foreign research on children's scientific thinking development and its revelation[J]. Western Quality Education,2022,8(05):122-124+136.DOI:10.16681/j.cnki.wcqe.202205038.

Luo Z S. Fundamentals of Item Response Theory[M]. Beijing: Beijing Normal University Press,2012: 1-4.

ANDRICH D. Controversy and the Rasch model: a characteristic of incompatible paradigms?[J]. Medical Care, 2004, 42（1）: 7-16

Qi S Q. Principles of Modern Education and Psychometrics [M]. Beijing: Higher Education Publishing Co. Beijing: Higher Education Publishing House, 2002: 91-96.

COE R. Comparability of GCSE examinations in different subjects: An application of the Rasch model[J]. Oxford Review of Education,2008, 34（5）: 609-636.

Li J L. A review of research on the application of Rasch model in China[J]. Journal of Guangdong University of Foreign Studies,2016,27(02):73-78.

Wu F T, Tian H ,Wang Y, Fan Y S. Research on the analysis of knowledge mastery and cognitive ability based on Rasch model under the vision of smart education[J]. Journal of East China Normal University(Education Science Edition),2021,39(08):57-69.DOI:10.16382/j.cnki.1000-5560.2021.08.005.

Bo Y, Zhu W Q, Chen H Z. The application of Rasch model in the quality analysis of test papers - An example of the sixth grade technology and engineering literacy assessment test paper of elementary science[J]. Educational Measurement and Evaluation,2019(01):25-31.DOI:10.16518/j.cnki.emae.2019.01.004.

Maja Planinic, Lana Ivanjek, Ana Susac. Rasch Model Based Analysis of the Force Concept Inventor [J].The American Physical Society, 2010, 3（10）: 1~11.

Wang L. The practice of Rasch objective isometrics in the PISA China pilot study [J]. Psychology Exploration, 2007,(4):69~73

Talib, A. M., Alomary, F. O., & Alwadi, H. F. (2018). Assessment of Student Performance for Course Examination Using Rasch Measurement Model: a Case Study of Information Technology Fundamentals Course. Education Research International, (3), 1-8.

Peng R Z, Zhu G C, Wu W P. Research on the quality of intercultural competence test based on Rasch model[J]. Foreign Language Community,2022,No.212(05):12-19+79.

Jan-Eric Gustafsson. Testing and obtaining fit of data to the Rasch model [J]. British Journal of Mathematical and Statistical Psychology, 1980（33）: 206~233.

Zheng Y, Zhang J P, Zhang Y F. Performance of high school students' scientific argumentation ability in physics - A test evaluation based on the Rasch model[J]. Physics Teacher,2019,40(01):2-6.

Bond T G & Fox C M. Applying the Rasch Model: Fundamental Measurement in the Human Sciences [M].Oxford: Psychology Press , 2013.

Engelhard Jr G. Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences.

Gao J B, Bai Y. A study on the evaluation of scientific literacy in the fourth grade of elementary school based on Rasch model[J]. Journal of Southeast University (Philosophy and Social Science Edition),2021,23(S1):135-138.DOI:10.13916/j.cnki.issn1671-511x.2021.s1.029