Disciplinary and Interdisciplinary Science Education Research

RESEARCH

Open Access

Development and validation of an instrument for assessing scientific literacy from junior to senior high school

Lina Zhang[1*], Xiufeng Liu[2] and Hua Feng[1]

Abstract

Assessing students' scientific literacy is necessary for science education policy, accountability and curriculum design and implementation, and there is a need for a valid, reliable and easy to use measurement instrument by local education authorities. Based on the definition and assessment framework of scientific literacy of PISA, a scientific literacy assessment instrument for junior high school and senior high school students was developed. The Rasch model was used to establish evidence for the validity and reliability of measures of the instrument. The results showed that the instrument measures met the validity and reliability expectations. Student scientific achievement levels were defined into four levels of Excellent, Advanced, Proficient and Basic using the Bookmarking method. The instrument can be used to assess student scientific literacy change from junior high school through high school to inform science education policy, accountability as well as curriculum design and implementation at local education levels.

Keywords Scientific literacy, Instrument Development, Validation, Assessment, Rasch modeling, Standard setting

Introduction

Scientific literacy is the competency to engage with science-related issues as a reflective citizen; it is the ability to meet complex needs by mobilizing psychological and social resources in specific contexts (OECD, 2006a, 2017). Scientific literacy involves multiple dimensions and has a multi-level hierarchical structure (Holbrook et al., 2009; Laugksch, 2000; Miller, 1983; Shen, 1975). Such a structure has also been described as a "rope" in which multidimensional and hierarchical interactions are integrated (Murcia, 2009). The description of the comprehensive performance of students to meet expectations of scientific literacy is the purpose of curriculum standards (e.g., NRC, 1996; 2012; MOE, 2022) and is the basis for the development of scientific literacy assessments.

Although the concept of scientific literacy has been put forward for many years and has been accepted as the goal of science education by many countries, how to best achieve scientific literacy for all students remains a challenge (Alonzo et al., 2012). One main impediment to achieve scientific literacy for all students is the lack of clear expectations of student performances. In the past, science curriculum standards in many countries usually emphasize the content standards without clearly stating how students should perform on the content standards. In recent years, performance standards on how students should demonstrate their competences have been explicitly included in science curriculum standards, such as the Next Generation Science Standards of the United States (NRC, 2012; NGSS Lead States, 2013), the Science Curriculum of the United Kingdom (DE, 2015), the Science Syllabus of Singapore (MOE, 2021), and Science Curriculum Standards in China (MOE, 2022). Performance expectations in the curriculum standards provide

*Correspondence:
Lina Zhang
zln000407@163.com
[1] Mathematics and Science Education Faculty, Beijing Institute of Education, No. 2, WenXing St., Xi Cheng District, Beijing 100044, People's Republic of China
[2] Department of Learning and Instruction, University at Buffalo, SUNY, 518 Baldy Hall, Buffalo, NY 14260-1000, USA

guidance for the development of assessment instruments of students' scientific literacy, which is currently a major task for many countries such as USA, UK, and China.

Large-scale international comparison studies in science typically define student performances based on international consensus expectations. For instance, TIMSS (Trends in International Math and Science Study) defines students' performance into four levels: Advanced, High, Intermediate and Low (Mullis et al., 2020). However, the performance standards in international science assessments such as the above TIMSS performance levels are not aligned with science curriculum standards of any country, thus they do not provide guidance to the implementation of science curriculum standards in any country. Also, these international large-scale assessments are complex in design (e.g., block design and matrix sampling) and time consuming; they are not practical to be used for monitoring the development of students' scientific literacy at local levels such as states/provinces and school districts. Low-stakes, valid and reliable instruments for assessing student scientific literacy by local education governments and authorities are urgently needed.

In this paper, we report an approach to developing scientific literacy assessments at local education levels. We developed a measurement instrument of scientific literacy for junior and senior high school students based on PISA scientific literacy conceptual framework. The instrument was intended to assess and monitor the trends of students' scientific literacy across secondary grades in order to improve the implementations of the national science education standards. The assessment results can be used to optimize local education policies, develop teaching resources and provide pertinent teacher professional development. In this paper, we report our effort to develop and validate this measurement instrument. The following research questions were answered:

(1) What is the evidence to support the validity and reliability claims of measures of the instrument for assessing scientific literacy for junior high and high school students?
(2) What are the different levels of students' performance in scientific literacy for junior high and senior high school students?

The first research question focused on the development and validation of the instrument; the second question focused on using the validated instrument to decide student scientific literacy performance levels which requires setting performance standards. In answering the above research questions, we contribute to the current science education literature by operationalizing scientific literacy based on a national science education standard; we also contribute to promoting scientific literacy by making available a valid and reliable measure of student scientific literacy across secondary school grades with clearly defined performance levels. The instruments can be used by local education authorities such as state, provincial department of education and school districts to monitor student achievement of scientific literacy so that timely and pertinent polices, resources and teacher professional development programs at local education levels may be implemented.

Literature review

Since the concept of scientific literacy was first put forward in the 1950s (Hurd, 1958), it has been adopted by many countries and regarded as the goal of science education (Reiss et al., 1999; NRC, 1996, 2012, 2015; Roberts, 2007). The international large-scale assessments of scientific literacy have also been developed and implemented; those assessment results have influenced national education policies and future directions. For example, the Organization for Economic Co-operation and Development (OECD) introduced an assessment called the Program for International Student Assessment (PISA) that includes an assessment of scientific literacy (OECD, 1999, 2006a, 2017). Results of PISA have had an extensive impact on the education policies around the world (Halász & Michel., 2011; Michel, 2017). Half of European countries explicitly referenced PISA (Waddington et al., 2007) when discussing science education standards. PISA and its results also have had a great impact on science education in China on assessment, teaching, educational policies and curriculum design (Lu, 2013; Xu, 2018).

A common interest among countries participating in international science assessment programs like PISA and TIMSS is to use the test results and other contextual data collected in the periodic surveys to identify factors that may be contributing to students'performance outcomes (Britton, et al., 2014; Zhai et al., 2023). Another use of results of the international science assessment is international comparisons among participating countries. A third use is monitoring of student science achievement over time. The use to monitor student science achievement over time are also the purpose of periodic national and state-level science assessment. For example, in the US NAEP is given to students of 4^{th}, 8^{th} and 12^{th} grades every four years for monitoring student achievement in subjects such as science at the national, state, and selected school district levels. NAEP reports (known as The Nation's Report Card) compare student performance in a given subject across states, within the subject over time, and among groups of students within the same grade.

One key feature of large scale science assessments is structured item development from a framework. For example, PISA has been designed to assess scientific literacy based on a common framework, not science curriculum of any country. PISA assessment framework emphasizes societal needs, especially the essential skills that future lives need, i.e., the competence of young adults in meeting literacy challenges of the future, including being able to analyze, reason, and communicate science-related ideas effectively and to continue learning throughout life. PISA adopts a multiple component scientific literacy framework, including scientific competencies, scientific knowledge, and context (OECD, 2017). The newly released PISA 2025 framework adds scientific attitudes/scientific identity as a new component of scientific literacy (OECD, 2023). Among them, competencies are the core of scientific literacy; knowledge and attitudes/scientific identity contribute to students' competencies. The scientific literacy embodied in real contexts requires people to use their competencies. PISA uses real contexts to anchor a set of 3–6 items into an item set/bundle.

Different from PISA, the purpose of TIMSS is assessing trends in science achievement over years. TIMSS seeks to examine students'achievement in terms of subject-matter knowledge and cognitive ability. TIMSS is curriculum oriented, thus its sampling is directly associated with grade levels in school. Specifically, TIMSS samples students in grades 4 and 8. On the other hand, PISA randomly samples 15-year-olds from the participating countries without specifying curriculums used by students. At the national and state levels, the purpose of NAEP is to assess educational progress every four years. Therefore, its sampling is at grades 4, 8 and 12. These grades represent education stages of elementary school junior high school, and high school. NAEP content is not tied to any state curricula.

Another key feature of large scale science assessments is precise specification of performance levels. For example, for NAEP in the US, National Assessment Governing Boady (NAGB) developed three levels of achievement for the assessment. *Basic* denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade. *Proficient* represents solid academic performance for each grade assessed. Students reaching this level demonstrate subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter. *Advanced* signifies superior performance. These levels are the primary means of reporting NAEP results to the general public and policy makers regarding what students should know and be able to do on NAEP assessments. For the purpose of monitoring trends in science achievement, TIMSS have also developed levels of achievement. There are four levels as international benchmark: low, intermediate, high and advanced. On the other hand, PISA have developed Level 1(including 1a, 1b) to Level 6 as proficiency levels of students' scientific literacy. Performance levels defined by NAEP and TIMSS are more meaningful than that of PISA for local governments, teachers and the public to understand and use.

Conceptual framework

PISA defined scientific literacy as the competence to meet complex societal needs by using and mobilizing psychosocial resources (including skills and attitudes) in specific contexts, and its constituent elements include scientific knowledge, scientific competency, scientific contexts, and cognitive demands (OECD, 2006a, 2017; Zhang, 2016). PISA emphasized understanding big ideas, inquiring and applying knowledge and skills in contexts, which are aligned with the Chinese science education standards (Lu, 2013; MOE, 2022; OECD, 2006a, 2017).

We adapted PISA scientific literacy framework to develop our assessment framework (see Fig. 1). Our assessment framework includes the following four aspects: 1) Scientific Knowledge; 2) Scientific Competencies defined by three aspects: Explain phenomena scientifically; Evaluate and design scientific inquiry; *and* Interpret data and evidence scientifically; 3) Scientific Context defined as natural and social science contexts in which students apply their scientific competencies; and 4) Cognitive Demand defined by the cognitive levels that need to be called to complete a task. In the above framework, scientific competency is the core of scientific literacy, and scientific knowledge is the foundation to form scientific competencies. These two dimensions are the primary aspects of assessment framework. Scientific context is a platform for students to demonstrate scientific competencies, and cognitive demand is used to define the difficulty of performing tasks.

In this study, the assessment instrument for scientific literacy of grades 6, 9 and 12 students was developed. The specific definitions of dimensions of each of the aspects are shown in Tables 1, 2 and 3. Specifically, Table 1 shows the assessment indicators of knowledge; Table 2 shows the assessment indicators of competence; and Table 3 shows the specific content of the situation/context. Tables 1, 2 and 3 are directly based on the scientific literacy framework defined by PISA 2015 (OECD, 2006a, 2017).

In terms of cognitive demands, the High level refers to analyzing complex information or data, integrating or assessment evidence, judging different sources of reasoning, and formulating a plan or a series of steps to solve

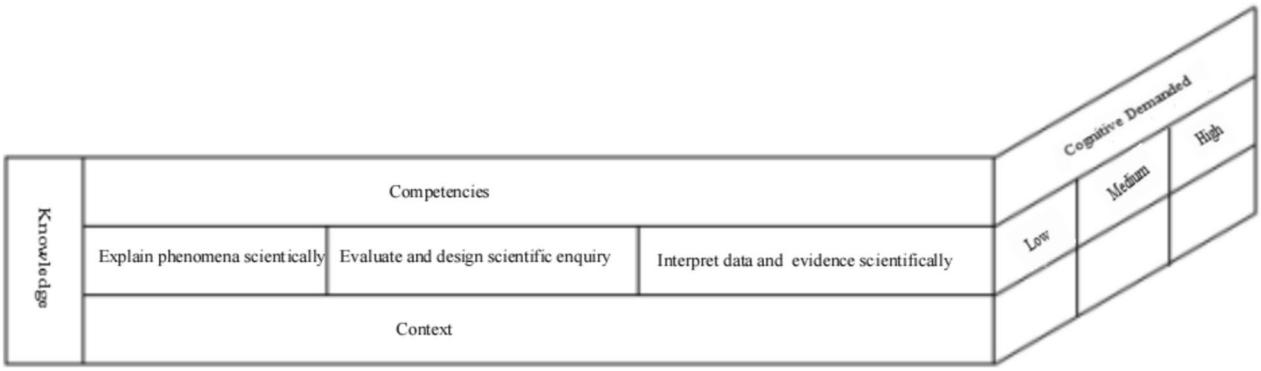

Fig. 1 Scientific literacy assessment framework. Adapted from Organization for Economic Cooperation and Development (2017)."PISA 2015 Science Framework", in PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving , OECD Publishing, Paris: 41

problems; the Medium level refers to the use and application of concepts to describe or explain phenomena, the selection of procedures involving two or more steps, the organization/presentation of data, or the interpretation of simple data sets or graphs; and the Low level refers to a one-step process of applying a fact, term, principle or concept, or locating single point information from charts and tables. This framework of cognitive demands followed the study of depth-of-knowledge taxonomy (Webb, 1997). Previous study of PISA (OECD, 2017) found that the four factors determined the cognitive demand of items: the number and degree of complexity of elements of knowledge demanded by the item; the level of familiarity and prior knowledge that students may have; the cognitive operation required by the item; and the extent to which forming a response is dependent on models or abstract scientific ideas. This four-factor approach allows for a broader measure of scientific literacy across a wider range of student ability. Categorizing the cognitive processes required for the competencies that form the basis of scientific literacy together with a consideration of the depth of knowledge required offers a model for assessing the level of demand of individual items. In addition, the relative simplicity of the approach offers a way to minimize the problems encountered in applying such frameworks for mapping items against the two dimensions of knowledge and competencies. In addition, each item can also be mapped using a third dimension based on a depth-of-knowledge taxonomy. This provides a means of operationalizing cognitive demand as each item can be categorized as making demands that are High, Medium, and Low. This cognitive demand framework is accepted in the current study.

Large-scale assessments have typically divided students' performance into 3–4 levels for local governments, teachers and the public to understand and use (Zhai et al., 2023). For example, TIMSS defines students' scientific achievements into four levels: excellent, advanced, proficient, and basic (Mullis et al., 2020). This performance standard was adopted by the current study. Specifically, student overall scientific literacy performance is defined into four levels: Excellent, Advanced, Proficient and Basic. The definition of the four performance levels is presented in Table 4.

Although we adapted PISA scientific literacy framework for our assessment, there are a few major differences between our framework and the PISA framework. PISA only assesses 15-year-old students' scientific literacy; we assess 6^{th}, 9^{th} and 12^{th} graders in order to monitor changes in student scientific literacy from beginning of junior high (i.e., end of elementary) to end of middle school to end of high school. In majority schools of China, grade 6 is the final year of elementary school, grade 9 is the final year of junior high school, and grade 12 is the final year of high school. Choosing these grades allows for monitoring the change in student scientific literacy at those key education stages. Different from PISA, our assessment aims for monitoring the development of scientific literacy over years to inform science education policy, accountability as well as curriculum design and implementation at local education levels. In addition, while PISA adopts six performance levels from 1a/1b to 6, our assessment adopts four performance levels from Basic to Excellent to be more meaningful to Chinese teachers and government officials.

Method

Item development

Previous research reported a method to develop instruments assessing elementary to high school students' growth in understanding the concept of matter based on the Rasch model (Liu, 2007). In order for student scores

Table 1 Scientific knowledge dimensions

Dimension	Knowledge
Content knowledge	Physics: Motion and forces (e.g. velocity, friction) and action at a distance (e.g. magnetic, gravitational and electrostatic forces) Energy and its transformation (e.g. conservation, dissipation) Interactions between energy and matter (e.g. light and radio waves, sound and seismic waves) Chemistry: Structure and properties of matter (e.g. particle model, bonds,,changes of state, thermal and electrical conductivity) Chemical changes (e.g. chemical reactions, energy transfer in chemical reaction) Energy and its transformation (e.g. chemical reactions) Biology: Cells (e.g. structures and function, DNA, plant and animal) The concept of an organism (e.g. unicellular and multicellular) Humans (e.g. health, nutrition, subsystems such as digestion, respiration, circulation, excretion,reproduction and their relationship) Populations (e.g. species, evolution, biodiversity, genetic variation) Ecosystems (e.g. food chains, matter and energy flow) Biosphere (e.g. ecosystem services, sustainability) Geography: Structures of the Earth systems (e.g. lithosphere, atmosphere, hydrosphere) Energy in the Earth systems (e.g. sources, global climate) Change in Earth systems (e.g. plate tectonics, geochemical cycles, constructive and destructive forces) Earth's history (e.g. fossils, origin and evolution) Earth in space (e.g. gravity, solar systems, galaxies) The history and scale of the universe and its history (e.g. light year, Big Bang theory)
Procedure knowledge	The concept of variables, including dependent, independent and control variables Concepts of measurement, e.g. quantitative (measurements), qualitative (observations), the use of a scale, categorical and continuous variables Ways of assessing and eneraliza uncertainty, such as repeating and averaging measurements Mechanisms to ensure the replicability (closeness of agreement between repeated measures of the same quantity) and accuracy of data (the closeness of agreement between a measured quantity and a true value of the measure) Common ways of abstracting and representing data using tables, graphs and charts, and using them appropriately The control-of-variables strategy and its role in experimental design or the use of eneraliza controlled trials to avoid confounded findings and identify possible causal mechanisms The nature of an appropriate design for a given scientific question, e.g. experimental, field-based or pattern-seeking
Epistemic knowledge	The constructs and defining features of science. That is: The nature of scientific observations, facts, hypotheses, models and theories The purpose and goals of science (to produce explanations of the natural world) as distinguished from technology (to produce an optimal solution to human need), and what constitutes a scientific or technological question and appropriate data The values of science, e.g. a commitment to publication, objectivity and the elimination of bias The nature of reasoning used in science, e.g. deductive, inductive, inference to the best explanation (abductive), analogical, and model-based The role of these constructs and features in justifying the knowledge produced by science. That is: How scientific claims are supported by data and reasoning in science The function of different forms of empirical enquiry in establishing knowledge, their goal (to test explanatory hypotheses or identify patterns) and their design (observation, controlled experiments, correlational studies) How measurement error affects the degree of confidence in scientific knowledge The use and role of physical, system and abstract models and their limits The role of collaboration and critique, and how peer review helps to establish confidence in scientific claims The role of scientific knowledge, along with other forms of knowledge, in identifying and addressing societal and technological issues

of different grades directly comparable, in this study we used about 1/3 of total questions in each assessment form for a grade level as common questions for two adjacent forms, i.e. elementary and junior high, and junior high and high school. We also used the single calibration to place student Rasch scale scores of different grades on a same scale.

According to the above described assessment conceptual framework, items were developed to assess student performances at defined different levels of scientific literacy. First of all, items were developed in bundles or groups for different scenarios, similar to those of PISA. Secondly, according to the development process of measurement instruments (Liu, 2020; Wilson, 2005), items were purposefully created for specific performance levels at different grades. Specifically, there were 6 item groups for Grade 6, with a total of 23 individual items and a full score of 46 points; there were 13 item groups for Grade

Table 2 Scientific competence dimensions

Dimension	Competencies
Explain phenomena scientifically	Recall and apply appropriate scientific knowledge Identify, use and generate explanatory models and representations Make and justify appropriate predictions Offer explanatory hypotheses Explain the potential implications of scientific knowledge for society
Evaluate and design scientific enquiry	Identify the item explored in a given scientific study Distinguish items that could be investigated scientifically Propose a way of exploring a given item scientifically Evaluate ways of exploring a given item scientifically Describe and evaluate how scientists ensure the reliability of data, and the objectivity and generalizability of explanations
Interpret data and evidence scientifically	Transform data from one representation to another Analyse and interpret data and draw appropriate conclusions Identify the assumptions, evidence and reasoning in science-related texts Distinguish between arguments that are based on scientific evidence and theory and those based on other considerations Evaluate scientific arguments and evidence from different sources (e.g. newspapers, the Internet, journals)

Table 3 Contexts in scientific literacy

	Personal	Local/National	Global
Health and disease	Maintenance of health, accidents, nutrition	Control of disease, social transmission, food choices, community health	Epidemics, spread of infectious diseases
Natural resources	Personal consumption of materials and energy	Maintenance of human populations, quality of life, security, production and distribution of food, energy supply	Renewable and non-renewable natural systems, population growth, sustainable use of species
Environmental quality	Environmentally friendly actions, use and disposal of materials and devices	Population distribution, disposal of waste, environmental impact	Biodiversity, ecological sustainability, control of pollution, production and loss of soil/biomass
Hazards	Risk assessments of lifestyle choices	Rapid changes (e.g. earthquakes, severe weather), slow and progressive changes (e.g. coastal erosion, sedimentation), risk assessment	Climate change, impact of modern communication
Frontiers of science and technology	Scientific aspects of hobbies, personal technology, music and sporting activities	New materials, devices and processes, genetic modifications, health technology, transport	Extinction of species, exploration of space, origin and structure of the universe

Table 4 Standards for students' performance levels of scientific literacy

Level	Description of level
Excellent	Students communicate understanding of concepts related to science in a variety of contexts
Advanced	Students apply understanding of scientific concepts
Proficient	Students show and apply some knowledge of science
Basic	Students show limited understanding of scientific principles and concepts and limited knowledge of science facts

Adapted from: Mullis, I., Michael, O.M. & Foy, P. et al. (Mullis et al. 2020)

9, with a total of 46 items and a full score of 92 points; there were 13 item groups in grade 12, with a total of 43 items and a full score of 86. Two same item groups, with a total of 8 individual items were included for both Grade 6 and Grade 9, and four same item groups,with a total of 13 individual items were included for both Grade 9 and Grade 12 to provide linkage in measures for test equating, as shown in Table 5. Item specification of scientific literacy assessment Instrument was shown in Additional File 1, Table 9. The proportion of linking items in Grade 6

Table 5 Item groups in scientific literacy assessment instrument

	Grade 6		Grade 9		Grade 12	
	Item group		Item group		Item group	
Physics	SS5	Sound	PJ1	"Practice 10"	PS1	"Magic Disc"
	L1	Friction	L1	Friction	PS2	Eelectromagnetic Induction Heating
			L3	Bus	L3	Bus
Chemistry	SS6	Change of Matter	CJ1	Food Additives	CS1	Orpiment and Realga
			L4	Graphene	L4	Graphene
					CS2	Acid Rain
Biology	SS3	Grassland Environment	BJ1	Food Chain	BS1	Clone Sheep
	SS7	Digestive Organs	BJ2	Cabbage Caterpillar	BS2	Parkinson's Syndrome
			BJ3	Myopia	BS3	Mouse Trap
			L5	Transgenic Technology	L5	Transgenic Technology
Geography	L2	Museum	L2	Museum	GS1	Haze
			GJ1	The Belt and Road Initiative	GS2	"Sponge City"
			L6	Earth's Revolution	L6	Earth's Revolution
			GJ2	Yellow River Basin		

J stands for junior high school, S stands for senior high school; SS stands for comprehensive science items, P stands for physics, C for chemistry, B for biology, and G for geography; L stands for the same item groups. For example, CJ stands for junior high school chemistry, CS stands for senior high school chemistry, L1 and L2 stand for the same two item groups in Grade 6 and 9 respectively: "friction" and "Museum"

instrument is 34.78%, which is 45.65%, 30.23% in Grade 9 instrument and Grade 12 instrument.

All items were scored by 0 ~ 2, that is, 2 points for full score, 1 point for partial score, 0 point for wrong or blank answer. Specifically, for items with one correct answer, students earn 2 points for selecting the correct answer, and 0 point for incorrect answers. For items with two correct answers, students who chose both the correct answers would get 2 points, 1 point for choosing either correct choice, and 0 point for incorrect answers. For items containing two parts, e.g., judging and explaining, students who answered both parts correctly would get 2 points, 1 point if correct only on one part, and 0 point for being incorrect on both parts.

All items were reviewed by five science education research experts and five science teacher experts to ensure content validity. Among the experts, there were one science education researcher and one science teacher of grade 6, one physics education researcher and one physics teacher of middle school, one chemistry education researcher and one chemistry teacher, one biology education researcher and one biology teacher, one geography education researcher and one geography teacher. The science education researcher and science teacher reviewed the 6th grade science items separately. They discussed different views and revised the items together, eventually reaching a consensus. Similarly, physics education researchers and physics teachers reviewed and revised physics items in grades 9 and 12, so did experts in chemistry, biology, and geography. For more information on item groups and how each item in each item group is aligned with four aspects: scientific knowledge, scientific competencies, scientific context and cognitive demand, please see the online Supplementary documents.

Items went through two rounds of revisions before they were given to a large sample of students for validation. The first round involved teacher reviews. Five 6^{th} grade science teachers completed the instrument of 6^{th} grade; five 9^{th} grade physics teachers, five 9^{th} grade chemistry teachers, five 9^{th} grade biology teachers, and five 9^{th} grade geography teachers responded to the questions of their subject areas of the instrument of 9^{th} grade. Finally, five 12^{th} grade physics teachers, five 12^{th} grade chemistry teachers, five 12^{th} grade biology teachers, five 12^{th} grade geography teachers responded to the questions of their subject areas of the instrument of 12^{th} grade. The teachers provided comments and suggestions to improve items. Items were revised accordingly. The second round revision was based on a pilot test involving students. We selected 90 students in each of grade 6, grade9 and grade12 respectively and asked them to complete the instrument. Descriptive statistical analyses (e.g., mean, standard deviation, difficulty, response distribution) were conducted and items were revised again.

Participants

A stratified purposeful sampling method was used to select schools and students involved in the scientific

literacy assessment in Beijing. First, students from 12 secondary schools in 6 districts (including 3 districts with good educational qualities and 3 districts with lower educational qualities) were selected to participate in the assessment. Second, the researchers selected students of excellent, medium and underdeveloped levels based on school-provided grades to take the assessment, which resulted in 1128 students from grades 6, 9 and 12, including 294 students from grade 6, 429 students from grade 9 and 405 students from grade 12.

Among them, there were 26.1% from 6^{th} grade, 38.0% from 9^{th} grade, and 35.9% from 12 grade.

Ethics clearance for the study was obtained from each school's School Oversight Board. The testing process was approved by the school leaders and science teachers, and the entire testing process was supervised by the teachers. All participants were informed of the purpose and procedure of the test and were told that their participation was voluntary and their anonymity would be guaranteed. The students have all consented to participate in the scientific literacy test. The researchers collected data from the schools that participated in the testing within one week. The test for each school was completed within one day (to avoid any potential communication between students in schools). The test took one hour in Grade 6, 1.5 h in Grade 9 and Grade 12. Only the researchers knew the schools, classes and students that participated in the test and there was no interchange of information between schools. All tests were in Mandarin Chinese. For reference outside China, the instrument was translated from Mandarin to English by an associate professor who majored in chemistry education; the English translation was then reviewed by a professor who is a native English speaker. The English translation of the instrument is available in Additional File 2 of the online supplementary documents.

Data analysis

Partial credit Rasch modeling was used to establish validity and reliability evidence of measures, and Winsteps3.72 and SPSS26.0 statistical analysis software were used to complete the data analysis. Specifically, the following Rasch analyses were conducted and outputs were examined: (1) reliability; (2) person-item match (Wright map); (3) dimensionality (principal component analysis of standardized residuals); and (4) item fit.

After the above validation by Rasch modeling, we created item booklets based on the difficulty values of the items produced by Rasch analysis, and engaged teachers to set standards of performance for grades 6, 9, and 12 into four levels respectively: excellent, advanced, proficient, and basic. Specifically, the difficulty measures of the items and students' performances as Rasch scale scores were exported into an Excel spreadsheet for standard setting by the Bookmarking method (Cizek et al., 2004; Wang, 2014) separately for grades 6, 9 and 12 students. Bookmarking method uses the item difficulties to arrange items from easy to difficult, and sets "bookmarks" between two items to indicate students' performance levels by experts (Cizek et al., 2004). Because the study assessed students in grades 6, 9 and 12, and there were three forms of measurement instrument with one form for each grade, experts were chosen for standard setting for each grade. Specifically, experts for each grade were 12 experienced science teachers, among them 1/3 came from schools with better student achievements, 1/3 from schools with medium student achievements and 1/3 from schools with lower student achievements. The experts for Grade 9 and Grade 12 standard setting were composed of teachers of physics, chemistry, biology and geography, with 3 experts in each discipline. Amongst the three experts in each discipline, one was from a school with best student achievements, one from a school with medium student achievements and one from a school with lower student achievements.

Result

Evidence to support the validity and reliability claims

Student reliability (person reliability) was 0.87 and item reliability was 0.99. Student separation index (person separation) was 2.56 and item separation index was 11.54. The above reliabilities and separation indices met reliability requirements of standardized measurement.

Second, the Wright map, which shows the internal structure of items and suggests construct validity of measures, demonstrates that the distribution of students' abilities matched the difficulty range of the items. Figure 2 is the Wright map. The left side shows the students' abilities and the right side shows the difficulty of items. It can be seen from Fig. 2 that the difficulty of items covers the whole distribution of students' abilities; There were only two small gaps located between CS1-1 and CJ3-2, and between BS2-2 and BS3 respectively.

Third, the principal components analysis of standardized restudies to test for unidimensionality of measures was conducted and the result is shown in Fig. 3. A variance greater than or equal to 50% explained by the Rasch measures can be regarded as evidence that the scale is unidimensional (Linacre, 2011); unidimensionality can also be assumed if the second dimension (first contract) in standardized residuals has the strength of less than 2 items (in terms of eigenvalues) and the unexplained variance by the first contrast is less than 5% (Oonet al., 2011).

Factor analysis of standardized residuals showed that almost all 91 items, except for items A (GJ4-1) and a (CJ1-1), had a loading within the -0.4 to + 0.4 range; the

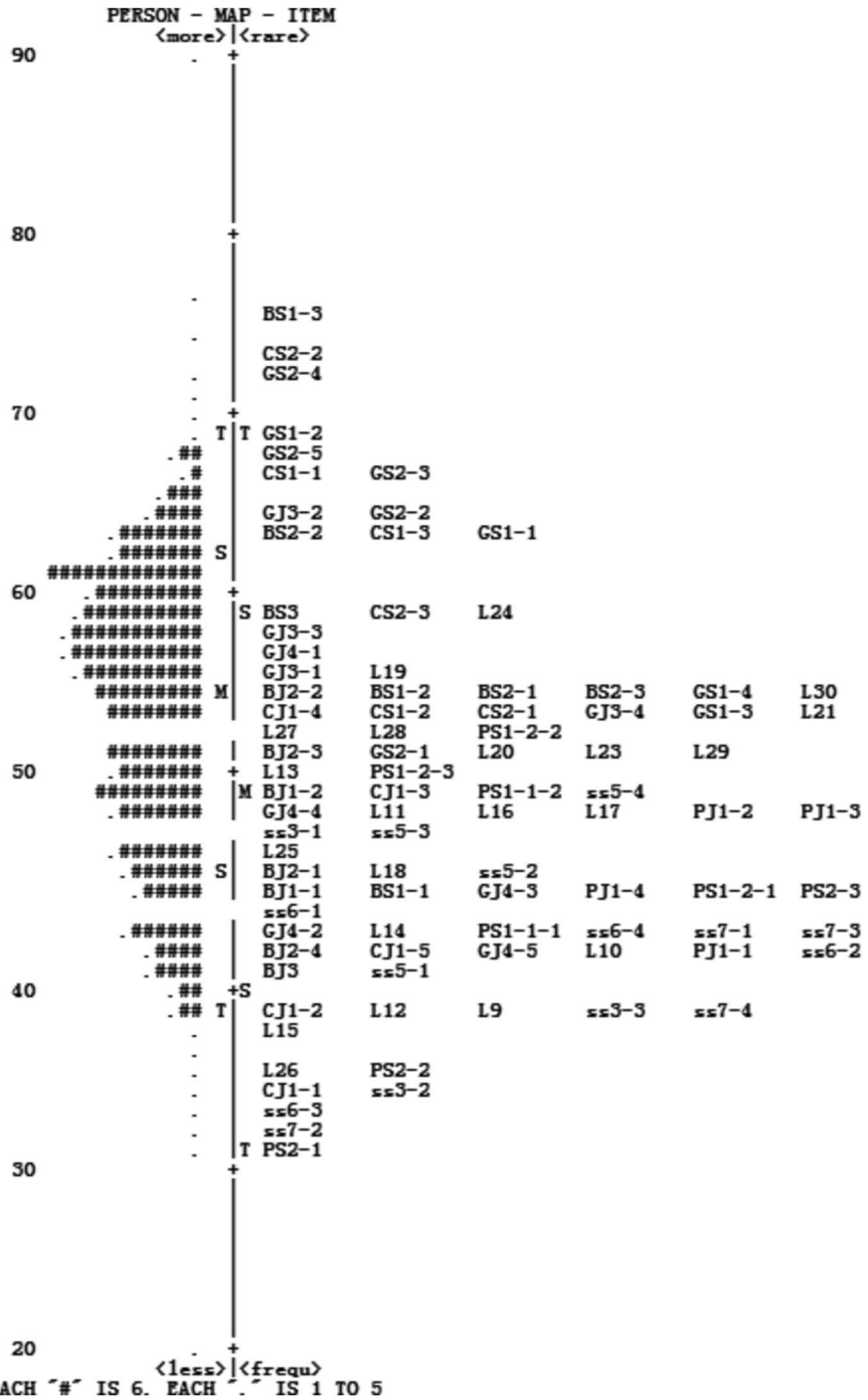

Fig. 2 Wright map for scientific literacy instrument

unexplained variance in the first contrast was 2.4 eigenvalues (1.8%, less than 5% of explained variance); only about 2 items possibly measuring more than one construct. However, the percentage of variance explained by Rasch measures was 32.5%, less than 50%. The above findings show that scientific literacy instrument measured a broad one-dimensional construct. Further investigation of the two items (GJ4-1 and CJ1-1) showed that GJ4-1 was a geography item, which tests students' competency to solve problems by applying information from text and pictures of the context. Students are required to understand information from the context of the "The Silk Road

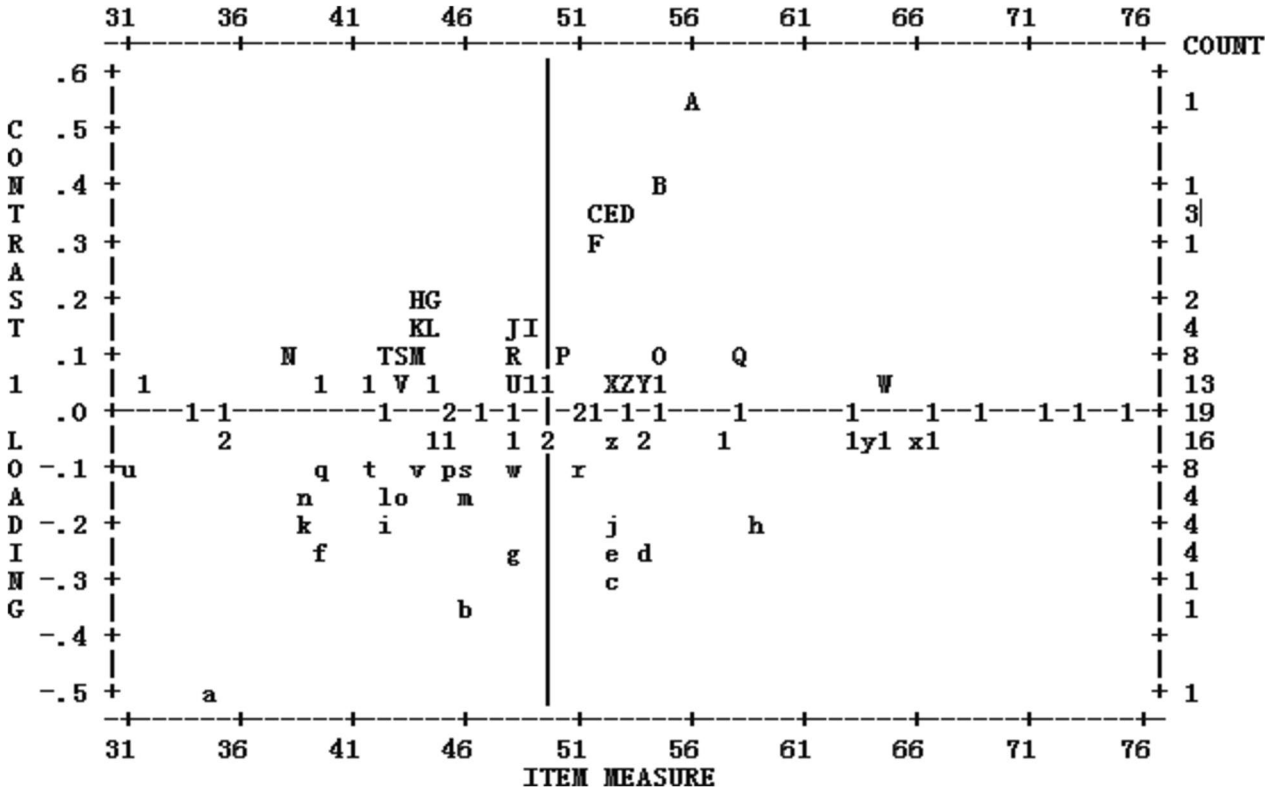

Fig. 3 Dimensionality for scientific literacy instrument

Belt" and apply geographical knowledge to write the continents crossed by the "The Silk Road Belt " and the dividing line between Asia and Europe. CJ1-1 is a chemistry item, which tests students' competency to extract information from a given context about why we use food additives. These two items measure competencies to obtain information from a given context and solve problems in science, but the difficulty of them were quite different: CJ1-1 only needs to extract information from the context, while GJ4-1 requires students to solve problems in combination with the geographical knowledge they have learned in addition to extracting information from the context. Considering the coverage of the knowledge, competencies, context and difficulty of the whole instrument, after discussion by experts, it was decided to retain these two items in the scientific literacy assessment instrument to maintain content validity.

Finally, the item fit was analyzed (see Table 10 in the Additional File 1 of the online supplementary documents). From the analysis of fit statistics of all the items, the standard error of measurement (SE) values of all 91 items were mostly less than 1 (SE value of 11 items were greater than 1). Using MNSQ between 0.7 and 1.3 as the standard, we can see that the Infit MNSQ values of all 91 items were within this range, and all Outfit MNSQ values of all items with the exception for PJ1-4, BJ2-3, GJ4-3, L26, L30, GS2-3 were also within this range. Taking the ZSTD value between-2 and 2 as the standard, we see that the Infit and Outfit ZSTD values of most items were within this range with exceptions for a few items such as SS3-1, SS6-2, L10, L13, L16, CJ1-4, BJ2-3. Given the large sample size, ZSTD statistics may not be given too much consideration. Based on the above findings, overall it can be considered that items of the measurement instrument had adequate fit with the Rasch model.

Standard setting of scientific literacy performance levels

Bookmarking method was used to set scientific literacy performance levels for grades 6, 9 and 12 students. A total of 36 expert science teachers participated the process. In each standard setting session, 12 experts in each grade read the item manual which arranged items from easy to difficult and set "bookmarks" between various items to indicate students' performance levels. Each expert set 3 bookmarks which indicate the cut-off point from basic to proficient, proficient to advanced, and advanced to excellent. After round 1 standard setting session, experts were provided descriptive statistics of the cut-off points they set for different performance levels and the percentages of students achieving each of the

levels. They then engaged in a discussion on their differences in decided cut-off points. Then the second round of standard setting took place independently. Although standard setting typically involves multiple rounds, e.g., at least three (Cizek et al., 2004), the number of standard setting rounds is determined by variation in standards set by experts. Because the variation in the standards set by experts was so small after round 2, the results of this second round were considered as the final standards. In a recent literature, two rounds of standard setting were also reported (Wang, 2014). Table 6 presents the descriptive statistics of the standards. The level below Proficient is considered Basic.

Table 6 shows three demarcation/cut-off scores of the three forms of the instrument from basic to excellent for grades 6 to 12. Based on the above 3 standards/cut-off scores in each grade, the 4 levels of students' scientific literacy were set. For grade 6, the demarcation score from basic to proficient was 39.75; the demarcation score from proficient to advanced was 45.68; and the demarcation score from advanced to excellent was 49.49. Therefore, students' scale scores of 39.75 or less belong to the Basic level, the scale scores of Proficient level are from 39.75 ~ 45.68, the scale scores of Advance level are from 45.68 ~ 49.49, and the scale scores of 49.49 or above belong to Excellent level. For Grade 9, the demarcation score from basic to proficient was 46.39; the demarcation score from proficient to advanced was 52.22; and the demarcation score from advanced to excellent was 58.33. Therefore, students' scale scores of 46.39 or less belong to the Basic level, the scale scores of Proficient level are from 46.39 ~ 52.22, the scale scores of Advance level are from 52.22 ~ 58.33, the scale scores of 58.33 or above belong to Excellent level. For Grade 12, the demarcation score was 50.71 from basic to proficiency, 58.09 from proficiency to advanced and 63.17 from advanced to excellent. Therefore, students' scale scores of 50.71 or less belong to the Basic level; the scale score of Proficient level are from 50.71 ~ 58.09, the scale scores of Advance level are from 58.09 ~ 63.17, the scale scores of 63.17 or above belong to Excellent level.

As can be seen from Table 6, standards for Proficient, Advanced and Excellent are different for Grades 6, 9 and 12, instead of one common set of standards for all grades. This is because curriculum standards and expectations for elementary, junior high and senior high school are different, and different standards for different grade levels allow more refined differentiation of student performances within each stage (i.e., elementary, junior high and senior high) for making decisions on instructional improvement and teacher professional development. Because of linking items between different forms and simultaneous calibration, student scores at different grade levels can still be directly compared by using the raw score to Rasch scale score conversion table (Table 8 below) and standards in Table 6.

Based on the above standards, the distribution of students' scientific literacy levels for students who participated in this study from grade 6 to grade 12 is presented in Table 7 and Fig. 4. We can see that the distributions of students of grade 6, grade 9 and grade 12 at the Basic level (Fig. 4, LEVEL 1) were 17.70%, 7.70% and 2.20% respectively. The distributions of students at the Proficient level were 37.10%, 27.30% and 30.10% (Fig. 4, LEVEL 2); the distributions of students at the Advanced level were 28.20%, 45.90% and 45.70% respectively (Fig. 4, LEVEL 3); and the distributions of

Table 7 Level distribution of scientific literacy of students in grades 6–12 ($N = 1\,128$)

Level	Grade 6	Grade 9	Grade12
Basic	17.70%	7.70%	2.20%
Proficient	37.10%	27.30%	30.10%
Advanced	28.20%	45.90%	45.70%
Excellence	17.00%	19.10%	22.00%

Table 6 Descriptive statistics of students' scientific literacy assessment standards

	Grade 6			Grade 9			Grade 12		
	Proficient	Advanced	Excellent	Proficient	Advanced	Excellent	Proficient	Advanced	Excellent
Mean	39.75	45.68	49.49	46.39	52.22	58.33	50.71	58.09	63.17
Median	39.58	45.68	49.60	46.35	52.21	58.86	50.88	58.02	63.15
Mode	39.58	45.68	49.60	46.35	52.21	58.86	50.88	58.02	63.15
Variance	0.47	0.64	0.13	0.02	0.00	1.51	0.33	0.06	0.01
Range	2.59	3.26	1.27	0.42	0.02	3.16	2.00	0.84	0.37
Minimum	39.33	44.61	48.33	46.35	52.21	55.70	48.88	58.02	62.97
Maximum	41.92	47.87	49.60	46.77	52.23	58.86	50.88	58.86	63.34

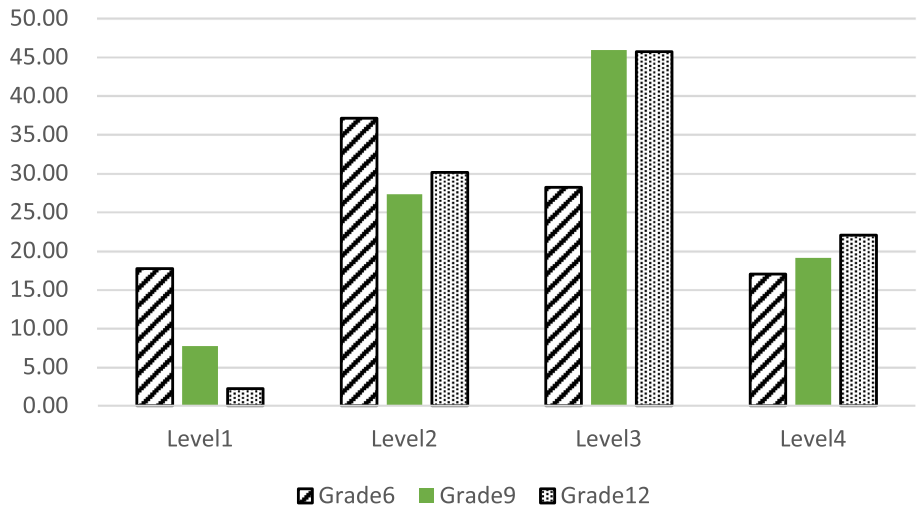

Fig. 4 Scientific literacy level distribution of students in grades 6–12

Table 8 Score conversion between Rasch scale scores and raw scores for different performance levels

level	Grade6		Grade9		Grade12	
	scale score	raw score	scale score	raw score	scale score	raw score
Basic	< = 39.75	< = 14	< = 46.39	< = 38	< = 50.71	< = 37
Proficient	39.75 ~ 45.68	14 ~ 24	46.39 ~ 52.22	38 ~ 55	50.71 ~ 58.09	37 ~ 45
Advanced	45.68 ~ 49.49	24 ~ 31	52.22 ~ 58.33	55 ~ 72	58.09 ~ 63.17	45 ~ 50
Excellence	> 49.49	> 31	> 58.33	> 72	> 63.17	> 50

students at the Excellent level were 17.00%, 19.10% and 22.00% respectively (Fig. 4, LEVEL 4).

Conclusion

This study defined a framework for scientific literacy assessment following the PISA framework and based on the Chinese national science education standards, and developed a measurement instrument. Rasch modeling was used to establish evidence for the validity and reliability claims. The results show that there was sufficient evidence for the reliability and validity claims of measures. Using the Bookmarking method to determine the demarcation scores, we also set performance standards for student performance at each grade and obtained the distributions of students' scientific literacy levels among the sample students.

Discussion

Regarding the assessment of scientific literacy, PISA defined the framework and assessment indicators of scientific literacy from four components: scientific knowledge, scientific competence, scientific attitude/identity and scientific context (OECD, 2006a, 2017, 2023). The scientific attitude/identity component is new to the 2025 PISA framework. This study followed the above framework. However, since the science curriculum of 15- to 18-year old students in Chinese Mainland implements discipline specific science curricula, the PISA assessment indicators of scientific knowledge are not entirely applicable to the actual curriculum implementation. The current study defined the assessment indicators of scientific literacy by disciplines according to the actual implemented curriculum. The revision was made to the PISA scientific knowledge assessment indicators, which is divided to different content of disciplines, i.e., physics, chemistry, biology and geography.

In order to incorporate contexts into items, PISA pioneered a type of item format which used a paragraph of text and1 to 2 pictures or charts to present scientific contexts to anchor a set of 3–6 items into an item group. This study adopted this item form, including item groups and scoring criteria (see Additional File 2 in the online supplementary documents). This item format may pose a challenge to Rasch modeling. That is, Rasch models require that items are independent from each other, the so-called local independence (Liu, 2020). Dependence between items undermines unidimensionality, which creates a major threat to the

validity of measures. According to the results of unidimensionality analysis reported above, measures of our instrument could be considered overall unidimensional, thus the local independence threat was not a concern. However, the fit statistics and the Wright map indicated that a few items could be further improved and a few additional items could be added in order to further improve the unidimensionality, thus the overall quality of the instrument.

As for the performance of students' scientific literacy, previous studies have divided students' scientific literacy into different levels according to assessment results (Mullis et al., 2020, OECD, 2006b, 2017; NAGB, 2019). The four levels of Basic, Proficient, Advanced and Excellent used in this study are based on the convention of Chinese schools. In China, schools and teachers are used to divide students into the four levels, such as excellent, advanced, meet the standard and fail; the four levels used during standard setting in this study approximate the conventional four levels with Basic equivalent to Fail and Proficient equivalent to Meeting the standard.

This study developed and validated an instrument for assessment of scientific literacy from junior high to high school. This instrument allows reporting of the distribution of scientific literacy levels for students in grades 6, 9, and 12. The change in distribution of levels in different grades may suggest a change in the quality of education. For example, a decrease in the proportion of high-level students (such as 'Excellent') and an increase in the proportion of low-level students (such as 'Basic') may indicate a decline in the quality of education, which calls for inquiry into reasons for such a change. Therefore, the scientific literacy assessment instruments developed and validated in this study can be used to monitor the change in scientific literacy of students at local levels for grades 6–12, providing information for science education policies, accountability, and curriculum reform.

In practice, the method used by teachers to set standards is based on students' raw scores of 60, 70, 80 and 90 points to divide four levels, which is not scientifically accurate (Wang, 2014). Bookmarking method is more scientific because it is based on an item order from easiest to most difficult according to the difficulty of the items; standard setting experts set standards based on what items are required in order for students to master in terms of scientific literacy. Thus, bookmarking method for standard setting does not depend on the performance of students who participated in the study. As the result, the standards set by bookmarking are applicable to other student samples.

Implications

The scientific literacy assessment instrument developed and validity in this study can be used to monitor students' scientific literacy across grades 6–12 in China, informing science education policy, accountability and curriculum design and implementation. The instrument can also be used by science teachers during the academic year to monitor students' learning, or carry out summative assessment such as grading.

In order to use the scientific literacy assessment instrument, a raw score to Rasch scale score conversion was needed (Liu, 2007). The Rasch scale scores and the corresponding original raw scores conversion was created by linear regression analysis. Three conversion equations below were established through linear regression for grade 6, grade 9 and grade 12 respectively:

Y = 1. 686X-52. 73 (1)	For Grade 6
Y = 2. 901X-96. 954 (2)	For Grade 9
Y = 1. 022X-14. 594 (3)	For Grade 12

Where Y is the Rasch scale score and X is the Raw score.

Based on the above three equations, a conversion table from the original raw scores (X) to the Rasch scale scores was also created (see Table 11 of Additional File 1 in the online supplementary documents). For example, the scale score of the sixth grade Basic level is 39.75, and the corresponding raw score is 14.29. Considering that each item in the instrument is scored 0, 1 and 2, the raw score of the sixth grade Basic level (14.29) should be rounded to 14. The scale scores of grade 6 Proficient level are from 39.75 ~ 45.68, and the corresponding raw scores are from 14.29 ~ 24.29, or rounded to 14 to 24. The scale scores of Advance level in Grade 6 are from 45.68 ~ 49.49, and the corresponding raw scores are 24.29 ~ 30.71, or rounded to 24 and 31. The scale score of Grade 6 Excellence level is 49.49, and the correspondent raw score is 30.71, or rounded to 31. Table 8 presents the conversion between Rasch scale scores and the corresponding raw scores for different performance levels at each grade level.

Limitation

One limitation of this study is the sample of students. Although this study adopted stratified purposeful sampling within the scope of Beijing, due to the difficulty of actual implementation, the proportion of schools with students of good performance in the ninth and twelfth grade is relatively high, which leads to the high distribution of advanced and excellence levels as shown in Table 8 and Fig. 4. Despite of this overall good performance, the range and order of item difficulties of the

measurement instrument should remain the same, an advantage of using Rasch modeling. This means that researchers and teachers in other provinces of China or other countries can use the instrument to measure student scientific literacy no matter how their students' scientific literacy may differ.

Another limitation of this study is that no Rasch analysis was conducted to analyze pilot testing data (we only conducted descriptive statistical analyses). Instrument development and validation typically need to go through multiple iterations and Rasch analyses. Given that the quality of items and the instrument as a whole are of high quality based on Rasch analysis results, this limitation may not prevent the instrument from being used or adapted by others, although further validation of the instrument would still be desirable.

Because the instrument was validated and performance standards were set using samples from Beijing, the instrument needs to be further validated and performance standards need to be set using samples from other countries following the method in this study. After student test, validation by Rasch modeling and bookmarking method should be conducted to set standards. Finally, the contexts of some item groups, especially those geography items, involve geography of China, they need to be revised when they are used to assess the scientific literacy of students from other countries.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43031-023-00093-2.

Additional file 1: Table 9. Item specification of scientific literacy assessment instrument. **Table 10.** Item-fit analysis of items. **Table 11.** Score conversion table between raw scores and Rasch scale scores.

Additional file 2. Scientific literary assessment instrument G6 science literacy test questions SS3 grassland environment.

Acknowledgements
Thanks for Ying Zhou, Fei Jin, Chunyan Li, Xiaoying Gao and their research group for data collection.

Authors' contributions
LZ and XL conceptualized the research, decide the methodology,performed the validation; LZ performed the data analysis and wrote the manuscript; LX supervised the research,reviewed and edited the manuscript; HF performed the validation and administrated the project. All authors read and approved the final manuscript.

Funding
Beijing Educational Science Planning Funding (BDAA2020029).

Availability of data and materials
The data underlying this article will be shared on reasonable request to the corresponding author.

Declarations

Ethics approval and consent to participate
Ethics clearance for the study was obtained from each school's School Oversight Board. The testing process was approved by the school leaders and science teachers, and the entire testing process was supervised by the teachers. All participants were informed of the purpose and procedure of the test and were told that their participation was voluntary and their anonymity would be guaranteed. The students have all consented to participate in the scientific literacy test.

Competing interests
Lina Zhang and Hua Feng have no conflicts of interest; Xiufeng Liu is co-editor-in-chief of DISER.

Received: 20 September 2023 Accepted: 29 November 2023
Published online: 11 December 2023

References

Alonzo, A. C., & Gotwals, A. W. (2012). *Learning Progressions in Science: Current Challenges and Future Directions*. Rotterdam, Boston, Taipei: Sense Publishers.

Britton, E. D. & Schneider, S. A. (2014). Large-scale assessments in science education. In S.K. Abell & N.G. Lederman (Eds.), *Handbook of Research in Science Education* (pp. 791–808). Routledge.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measures: Issues Practice, 23*(4), 31–50.

Department for Education (2015).National curriculum in England: science programmes of study. Retrieved November 18, 2022,from https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study.

Halász, G., & Michel, A. (2011). Key competences in Europe: Interpretation, policy formulation and implementation. *European Journal of Education, 46*(3), 289–306.

Holbrook, J., Rannikmae, M., Coll, R. K., & Taylor, N. (2009). The meaning of scientific literacy. *International Journal Environmental and Science Education, 4*(3), 275–288.

Hurd, P. D. (1958). Science literacy: its meaning for American schools. *Educational Leadership, 16*(1), 13–16.

Laugksch, R. C. (2000). Scientific literacy: a conceptual overview. *Science Education, 84*(1), 71–94.

Linacre, J. M. (2011). A user's guide to Winsteps: Rasch-model computer programs. http://winsteps.com.

Liu, X. (2007). Elementary to high school students' growth over an academic year in understanding concepts of matter. *Journal Chemical Education, 84*(11), 1853–1856.

Liu, X. (2020). *Using and developing measurement instruments in science education: A Rasch modeling approach* (2nd ed.). Information Age.

Lu, J. (2013). *The theory and practice of PISA*. Shanghai: East China Normal University Press.

Michel, A. (2017). The contribution of PISA to the convergence of education policies in Europe. *European Journal of Education, 52*(2), 206–216.

Miller, J. D. (1983). A conceptual and empirical review. *Daedalus, 112*(2), 29–48.

Ministry of Education (MOE) of Singapore.(2021).Science Syllabuses Lower Secondary: Express Course & Normal (Academic) Course. Retrieved November 18, 2022, from https://www.moe.gov.sg/-/media/files/secondary/syllabuses/science/2021-science-syllabus-lower-secondary.ashx?la=en&hash=21D677EC03ED15C456412AB2FCD2979579408CFD.

Ministry of Education (MOE) of PRC. (2022). *Full-time compulsory education science curriculum standard* (2022nd ed.). Beijing Normal University Press. (in Chinese).

Mullis, I., Michael, O. M., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. Chestnut Hill: TIMSS & PIRLS International Study Center, Lynch School of Education and Human

Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).
Murcia, K. (2009). Re thinking the development of scientific literacy through a rope metaphor. *Research in Science Education, 39*(2), 215–229.
National Assessment Governing Board (NAGB) of US. (2019). *Science framework for the 2019 national assessment educational progress*. National Assessment Governing Board.
National Research Council (NRC) of US. (1996). *National Science Education Standards*. National Research Council.
National Research Council (NRC) of US. (2012). *A Framework for K-12 science education: Practices, crosscutting concepts and core ideas*. The National Academies Press.
National Research Council (NRC) of US. (2015). *Guide to implementing the next generation science standard*. The National Academies Press.
NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. The National Academies Press.
Organization for Economic Co-operation and Development (OECD). (2006a). *Assessing scientific, reading and mathematical literacy, a framework for PISA 2006*. OECD Publishing. https://doi.org/10.1787/9789264026407-en
Organization for Economic Co-operation and Development (OECD). (1999). *Measuring student knowledge and skills: A new framework for assessment*. OECD Publishing.
Organization for Economic Co-operation and Development (OECD). (2017). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing.
Organization for Economic Co-operation and Development (OECD). (2006b). *PISA 2006 Technical Report*. Retrieved November 18, 2022, from https://www.oecd.org/pisa/data/42025182.pdf
Organization for Economic Co-operation and Development (OECD). (2023). *PISA 2025 Science framework (draft)*. https://pisa-framework.oecd.org/science-2025/assets/docs/PISA_2025_Science_Framework.pdf.
Oon, P. T. & Subramaniam, R. (2011). Rasch modeling of a scale that explores the take-up of physics among school students from the perspective of teachers. In B.J. Fraser & J. P. Dorman (Eds.), *Applications of Rasch measurement in learning environments research* (pp. 119–139). Sense Publishers.
Reiss, M. L., Millar, R., & Osborne, J. (1999). Beyond 2000: science education for the future. a report with ten recommendations. *Journal of Biological Education, 33*(2), 68–70.
Roberts, D.A. (2007). Scientific literacy/science literacy. In S.K. Abell & N.G.Lederman (Eds.), *Handbook of Research in Science Education* (pp. 729–780). Lawrence Erlbaum Associates.
Shen, B. S. (1975). Science literacy: Public understanding of science is becoming vitally needed in developing and industrialized countries alike. *American Scientist, 63*(3), 256–268.
Waddington, D., Nentwig, P., & Schanze, S. (2007). *Making it comparable: Standards in science education*. Munster, Germany: Waxmann.
Wang, X. (2014). The application of Bookmark method in setting demarcation scores in standard based education. *China Examination, 24*(7), 10–18. (in Chinese).
Webb, N. L. (1997). *"Criteria for alignment of expectations and assessments in mathematics and science education",*. Washington: Council of Chief State School Officers and National Institute for Science Education Research Monograph, National Institution for Science Education.
Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.
Xu, S. H., et al. (2018). *PISA2015 basic literacy research report on 15-year-old students in four provinces and cities of China*. Guangzhou: Guangdong Higher Education Press. (in Chinese).
Zhai, X. M., & Pellegrino, J. W. (2023). Large-scale assessments in science education. In N. G. Lederman, D. L. Zeidler, & J. S. Lederman (Eds.), *Handbook of Research in Science Education* (pp. 1045–1097). Routledge.
Zhang, L. N. (2016). Implementation of PISA2015 scientific literacy assessment on Chinese science teaching and assessment. *Global Education, 45*(3), 15–24. (in Chinese).

Publisher's Note