MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2023, 19(9), em2328
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/13524

Evaluating students' ability in constructing scientific explanations on chemical phenomena

Lukman Abdul Rauf Laliyo [1*], Rahmat Utina [2], Rustam Husain [2], Masri Kudrat Umar [2], Muhammad Rifai Katili [2], Citra Panigoro [3]

[1] Department of Chemistry Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Gorontalo, INDONESIA
[2] Postgraduate Program of Education Technology, Universitas Negeri Gorontalo, INDONESIA
[3] Department of Aquatic Resource Management, Faculty of Fisheries and Marine Science, Universitas Negeri Gorontalo, INDONESIA

Received 31 August 2022 ▪ Accepted 16 June 2023

Abstract

Evaluation of students' ability in constructing scientific explanations on scientific phenomena is essential as an effort to obtain information and feedback for innovation in learning process and curriculum development. Unfortunately, this issue is still left unexplored by researchers in chemistry education. Such is presumed to occur due to validated instruments, measurements, analysis techniques, and diverse epistemological values that leave much space to be investigated. Employing a Rasch model, we intended to validate test of ability in constructing scientific explanations on chemical phenomena, examine data fit with the Rasch model, evaluate difference in the students' ability in constructing scientific explanations, investigate items with different functions, and diagnose causes for difference in item difficulty level. The respondents were 550 students from seven senior high schools in three regencies/cities and 153 university students in Sulawesi, Indonesia. Data were collected by 30 test items; each item consisted of three questions measuring students' ability in proposing their knowledge (Q1), evidence (Q2), and reasoning (Q3). Their responses were assessed on criteria and analyzed using the Rasch partial credit model. This model applies an individual-centered statistical approach allowing researchers to measure up to item and individual level. Results suggested that data fit the Rasch model measurement. Also, students' ability in constructing scientific explanations varied significantly. We found no items with different functions, signifying that sex and hometown do not influence students' ability. However, based on item logit value grouping, it was discovered that item difficulty level also varied among students. This was particularly due to students' lack of chemistry concepts mastery that lowered their ability and accuracy in constructing scientific explanation. This shows lack of epistemological engagement of students in learning process. In conclusion, this study provides valuable insights into students' ability to construct scientific explanations and sheds light on factors that influence their performance in this area. Findings highlight need for targeted interventions that address students' conceptual understanding and engagement with chemistry concepts, as well as promote critical thinking and scientific reasoning skills. This has important implications for science education and can inform curriculum development and evaluation policies.

Keywords: evaluation, argumentation, explanation, scientific, phenomena, Rasch model

INTRODUCTION

The ability to construct scientific explanations has become a priority in future scientific argumentation development. Such an ability is one of the complex cognitive skills based on scientific facts requiring proper reasoning between theories and evidence, including critical thinking (Berland & Reiser, 2009; Mao et al.,

✉ lukman.laliyo019@gmail.com **(*Correspondence)** ✉ rahmatutina@gmail.com ✉ rustamhusain@gmail.com
✉ masriumar@gmail.com ✉ mrifaikatili@ung.ac.id ✉ citrapanigoro@ung.ac.id

Contribution to the literature

- This study contributes to the reinforcement of curriculum development and evaluation policies, especially how to evaluate students' MPI abilities, use a chemical phenomenon test instrument that is developed and validated by a Rasch modeling-based psychometric analysis technique.
- The results of this research can serve as evaluative information regarding the application of the 2013 chemistry curriculum of the chemistry subject in high schools in Indonesia.
- Also, this research offers new insight for teachers, researchers, and decision-makers in mapping out the students' ability in constructing scientific explanations through a scientific inquiry-based learning process.

2018). The scientific explanation contains a structure of two components:

(1) explanandum or the phenomenon to be explained and must be an indisputable fact and

(2) explanans or elements that make the facts in question understandable (Osborne & Patterson, 2011; Yao & Guo, 2018).

To construct scientific explanations effectively, students need to have a strong understanding and command of the scientific content. This knowledge allows them to apply their understanding to identify the key principles and characteristics that form the basis of scientific explanations for various phenomena (Wang, 2015). This means that scientific explanation is understood as the explicit application of theory that goes beyond the description of scientific patterns, either to reveal causal relationships or to model the mechanisms underlying certain situations or phenomena (National Research Council, 2012; Osborne & Patterson, 2011; Yao & Guo, 2018). From these perspectives, the involvement of students in constructing scientific explanations is pivotal as it can stimulate their creativity, problem-solving, and productivity in providing scientific elaboration, justification, and argumentation (Rahayu, 2019; Talanquer, 2018).

The construction of scientific explanations of chemical phenomena, such as rusting iron, melting ice, and the use of detergent, is an inseparable part of chemistry subject in high school. In this context, the benchmark of students' ability refers to explaining how or why a chemistry phenomenon occurs. Students are expected to utilize three main components, namely claim, evidence, and reasoning, to provide scientific responses to questions about chemical phenomena (Berland & Reiser, 2009; Chin & Brown, 2000; Jin et al., 2021). A claim is defined as the initial answer or explanation of a question in the form of short answers, including agreement, refutation, classification, grouping, or numbering. Evidence is examples or data given to support a claim. The examples or data can be based on the learning process, experience, experiment, or daily event. The reasoning is the elaboration that connects between claim and evidence. Sound reasoning should also refer to complementary knowledge, such as figures, graphs, and mathematical formulas (McNeill & Krajcik, 2008).

Some previous research has been carried out involving teachers and students in developing the skills of building, analyzing, and evaluating students' scientific explanations (Berland & Reiser, 2009; Wang, 2015). Research by Yao and Guo (2018) showed that the development of students' scientific explanations could be evaluated using a framework that includes four main components: phenomena, theories/claims, data/evidence, and reasoning. Yao and Guo (2018) concluded that variations in students' ability to construct scientific explanations could be attributed to disparities in learning conditions and practices Research conducted by Jin et al. (2021) focused on the development of an argumentation framework, which encompasses four levels: non-causal arguments, causal arguments with no logical connection, causal arguments with weak reasoning, and causal arguments with strong reasoning. This framework is particularly relevant to the construction of scientific explanations as argumentation plays a crucial role in providing logical reasoning and evidence to support scientific claims. By utilizing this framework, students can enhance their ability to construct scientifically sound explanations by employing effective argumentation strategies and strengthening their reasoning skills. The results show that these levels of argumentation can be empirically proven to be different, and the order of differences between the levels of argumentation is significant. These previous studies have shown that the involvement of students in the learning process and scientific experiment has distinct and epistemic characteristics (Ministry of Education and Culture, 2014; Deng & Wang, 2017; Driver et al., 2000; Erduran et al., 2004; Osborne et al., 2004). For the last few decades, experts have explored new methods and curricula to develop students' ability to construct scientific explanations through meaningful involvement in learning (Hong et al., 2013). Such a notion instigated scholarly explorations on evaluating students' ability to construct scientific knowledge (Mendonça & Justi, 2014; Sandoval & Millwood, 2005). However, the implementation of such a research topic in chemistry education is still lacking (Cetin, 2014; Rahayu, 2019).

In Indonesia, the development of the ability to construct scientific explanations has been mandated in the 2013 curriculum. For the last decade, this focus has

been applied explicitly through constructivist approach-based learning and scientific inquiry-based learning in classrooms and laboratories across all high schools in Indonesia (Rahayu, 2019). Unfortunately, based on literature searches, only a few studies evaluate the students' ability to construct scientific explanations in chemistry. Not many chemistry education researchers in Indonesia concentrate on the research topic in question. Thus, feedback and information regarding students' learning progress in constructing scientific explanations are minimal. This research gap could pose challenges in finding additional information regarding the efficacy of incorporating learning activities based on a constructive approach and scientific inquiry-based learning within the 2013 chemistry curriculum. This difficulty is anticipated due to the scarcity of validated instruments, constraints in measurement and analysis techniques employed, and the need for further exploration of the diverse epistemological values inherent in learning conditions and practices.

The Rasch model is a popular psychometric model used in educational research to measure the abilities of individuals and the difficulty of items in educational assessments. Its applications include test development, item analysis, test calibration, person measurement, test equating and scaling, and item banking. By employing the Rasch model, researchers and educators can enhance the quality and precision of educational assessments, measure individuals' abilities, establish common measurement scales, and create adaptive testing systems. The Rasch model offers valuable insights into the difficulty of test items and the abilities of individuals, enabling meaningful comparisons and tracking of progress over time. A recent literature review by Yang et al. (2021) examined the current trends in Rasch modeling in educational research. The literature review examined 225 articles published from 2016 to 2020 that utilized Rasch modeling in educational research. The analysis revealed widespread use of Rasch modeling in diverse educational fields, including language assessment, science education, and mathematics education. Multidimensional Rasch models measure multiple latent constructs simultaneously, providing a comprehensive perspective on individual abilities. In a recent study by Lu and Chen (2021), the Rasch model was employed to evaluate the psychometric properties of an online reading comprehension test for Chinese EFL learners. The study demonstrated the Rasch model's efficacy in assessing test validity, reliability, item bias, and individual differences in ability. Apart from its utility in constructing assessments (Lu & Chen, 2021), the Rasch model has proven to be valuable in analyzing data from preexisting surveys and assessments (Nongna et al., 2023). This demonstrates the versatile nature and wide-ranging applications of the Rasch model within the realm of educational research. A study conducted by Nongna et al. (2023) focused on using Rasch analysis to assess the performance of higher education instructors in a Thai public university. The study employs a design-based research method that involves four phases: analyzing the results of performance assessment, formulating standards-setting appraisal, applying trial and quality inspection, and improving the standards-setting appraisal approach. The findings of the study suggest that the standards-setting appraisal approach is relevant for use as a criterion for recruiting and selecting higher education instructors. However, the researchers note that the transition point for determining competency levels may be accurate and consistent for instructors with moderate to high competency levels but may not be suitable for evaluating those with low competency levels. The study highlights the importance of instructors possessing high core competencies to meet the demand for quality teaching in higher education. The results of the study may have implications for the recruitment and selection of instructors, as well as the development of sustainable human capital in the higher education setting. Overall, the literature indicates that the Rasch model is a valuable tool for educational researchers across various domains. It can provide insight into individual abilities, item difficulty, and test quality, leading to improved assessments and instructional practices.

There are several commonly used instruments in the development of chemistry educational materials. Firstly, chemical concepts inventory (CCI) that being developed by Mulford and Robinson (2002). The study developed an inventory to assess alternative conceptions among first-semester general chemistry students, and it specifically focused on evaluating students' misconceptions in chemistry (Mulford & Robinson, 2002). Secondly, the assessment and analysis of the psychometric properties of CCI developed by Barbera (2013). It is a widely used instrument for assessing students' conceptual understanding of fundamental chemistry concepts. It consists of multiple-choice questions that cover various topics in chemistry, including atomic structure, chemical bonding, and chemical reactions (Barbera, 2013). The next instrument is chemistry self-efficacy scale (CSES). This instrument assesses students' self-efficacy beliefs related to chemistry, which refers to their confidence in their ability to perform tasks and succeed in chemistry-related activities. It typically includes items that measure students' confidence in their ability to understand, apply, and solve chemistry problems (Uzuntiryaki & Aydin, 2009). The last instrument is called attitudes toward chemistry scale (ATCS). This instrument measures students' attitudes and perceptions towards chemistry, including their interest, enjoyment, and motivation to learn chemistry. It may include items related to students' perceptions of the relevance of chemistry, their confidence in their chemistry abilities, and their interest in pursuing further studies or careers

in chemistry (Cheung, 2009). However, it is important to note that the aforementioned chemistry instruments, including CCI, CSES, and ATCS, were not developed using the Rasch measurement model.

The instruments mentioned above differ from Rasch analysis in this study in two ways. Firstly, these instruments primarily consist of self-report questionnaires or tests designed to assess various aspects of students' or teachers' perceptions, beliefs, attitudes, or knowledge related to chemistry. They provide valuable insights into individuals' subjective experiences and perspectives. On the other hand, Rasch analysis is a statistical method used to examine the psychometric properties of test or questionnaire data, which involves analyzing factors such as item difficulty, person ability, and item fit within a unidimensional construct. It focuses on the measurement properties of the instrument itself and aims to establish a quantitative understanding of the relationships between items and respondents. Therefore, while the listed instruments capture subjective aspects through self-report measures, Rasch analysis complements them by providing objective measurements and insights into the psychometric properties of the instrument. Both approaches contribute valuable information to the overall understanding of chemistry education but serve different purposes in terms of data analysis and interpretation. Secondly, the instruments listed above typically use multiple-choice questions or Likert-type scale items, while Rasch analysis involves analyzing the responses to each item in relation to the overall performance of individuals on the test or questionnaire. Rasch analysis also allows for the examination of item fit statistics, item difficulty, person ability, and other psychometric properties, which may not be directly assessed by the instruments listed above (Chan et al., 2021; Chi et al., 2021, Sumintono & Widhiarso, 2014). Rasch has been used to develop a number of student oriented instruments in multiple areas of science such as computational thinking, biology, and physics (Malone et al., 2021; Salibašić Glamočić et al., 2021; Sovey et al., 2022).

The purposes of this research are, as follows:

(1) to validate the test of constructing scientific explanations on chemical phenomena by the Rasch model,

(2) to examine the data fit with the Rasch model,

(3) to evaluate the difference in the students' ability in constructing scientific explanations on chemical phenomena,

(4) to investigate items with different functions, and

(5) to diagnose the causes for the difference in item difficulty level.

Student responses were assessed based on the criteria and analyzed using the Rasch partial credit model (PCM) approach using the WINSTEPS version 4.5.5 software. The analysis adopted an individual-centered statistic approach that allows the measurement up to the individual level of each student and each item. Henceforth, the research questions are formulated, as follows:

RQ1. To what extent do the data collected using the instrument developed in this study fit the Rasch model?

RQ2. How do students' ability to construct scientific explanations on chemical phenomena vary across different classes, genders, and hometowns?

RQ3. Based on the measurement result in the item level, are there any items that have different functions with regard to different sex and hometown of the students?

RQ4. Based on the measure of item logit, in what ways does the students' difficulty level differ from each other in constructing scientific explanations on chemical phenomena?

RQ5. Based on the item response patterns, what causes the different difficulty levels of students in constructing scientific explanations about chemical phenomena?

METHOD

Research Design

The quantitative descriptive study relied on a non-experimental study in which the students' ability in constructing scientific explanations on chemical phenomena was treated as a variable measured by a multi-tier multiple choice test. Prior to conducting the research, it was ensured beforehand that students had experienced formal learning in high school based on the 2013 curriculum. No intervention was carried out in their learning experiences, both in the learning process and materials. In other words, no treatment whatsoever was given to students that allowed them to have the ability to complete the questions in the instrument. This means that students' ability to construct scientific explanations is solely obtained through formal learning using the 2013 curriculum. In this curriculum, the achievements of core and basic competencies have been described in detail, including the sequence of learning activities, strategies applied, media and evaluation. The goal is to help teachers develop adaptive instructional plans and strategies for the learning needs of students so that students get a deep understanding of a topic they are studying (Bailey et al., 2012). In addition, this curriculum has described how the teacher's role in developing learning can facilitate students to find knowledge from various learning sources with the help of technology and encourage students to be involved epistemologically in the activities of formulating

Table 1. Respondents' demographic profile (n=703)

Demography	Code	Total	
		n	%
Students' classes			
Students at Senior High School A	A	170	24.0
Students at Senior High School B	B	83	11.7
Students at Senior High School C	C	14	2.0
Students at Senior High School D	D	22	3.2
Students at Senior High School E	E	53	7.5
Students at Senior High School F	F	135	19.0
Students at Senior High School G	G	73	1.7
Chemistry college students (1st-year)	H	38	5.4
Chemistry college students (2nd-year)	I	58	8.3
Chemistry college students (3rd-year)	J	57	8.1
Sex			
Male	M	158	22.5
Female	F	545	77.5
Hometown			
Gorontalo	A	289	41.1
Limboto	B	188	26.7
Bolango	C	226	32.2

problems, conducting experiments, connecting data, providing scientific explanations, and constructing arguments (Grooms, 2020; Duran & Dokme, 2016; Szalay & Tóth, 2016; Wu & Hsieh, 2006). The application of this curriculum is expected to form students with creative and independent characteristics capable of thinking critically in building a scientific explanation of a phenomenon based on scientific facts and forming relationships based on evidence and logical reasoning (Berland & Reiser, 2009). In this context, teacher involvement in implementing formal inquiry learning in the classroom is essential to determining the 2013 curriculum's success. Unfortunately, so far, it tends to be difficult to find studies that specifically explain the success rate of implementing this curriculum.

In quantitative research, the output data typically consist of numerical values. Meanwhile, the scores were derived by categorizing students' responses on each item according to predefined criteria or categories that reflect their ability levels. This step involves using deductive reasoning to derive specific data collection approaches from abstract concepts, and then obtaining precise numerical information through those approaches. The numerical data obtained represent a standardized, compact, and unified technique that measures the abstract concept empirically. In other words, it involves translating abstract concepts into concrete data through a standardized and unified approach for empirical measurement (Neuman, 2014). The obtained data were further analyzed quantitatively using the partial credit Rasch model analysis (Chi et al., 2021, 2022).

Respondents

This research was conducted at the end of the even semester, January to June 2022, which means that ABCDEFGH respondents have experienced formal learning of the basic chemistry concepts according to the 2013 curriculum. Still, the present work did not discuss whether or not the formal learning process effectively develops students' conceptual mastery. This research focuses mainly on measuring the ability to construct scientific explanations and does not place learning effectiveness as the object being discussed. One of the examples is students' scientific explanation of the phenomenon of rusting iron. This phenomenon can be explained by students who have mastered the concept of redox reactions properly and correctly. Based on the 2013 curriculum, this concept is learned by tenth-grade students. The effectiveness of learning using the 2013 curriculum is a part that can be reflected and explained based on the results of measuring the abilities of the intended students. Thus, the results of measuring students' abilities in constructing scientific explanations about chemical phenomena can be used as a reflection to find out how far the success rate of the 2013 curriculum implementation process in Indonesia is. Concerning the principles and ethics of research, students who voluntarily participated in this research had been asked for their consent as per the regulations of the Institutional Review Board (IRB). Students' identity remains confidential, and all information is for scientific development (Taber, 2014). **Table 1** shows respondents' demographic profile.

Instrument Development and Procedures

The basis for developing measurement instruments includes three main components: cognition, observation, and interpretation. Cognition refers to a theoretical path or a construct that helps students develop their abilities in certain knowledge domains. Observation is the ability of students based on the type of assessment of specific tasks and situations. Interpretation is a statistical model depicting expected patterns determining students' skills (National Research Council, 2007, 2012).

From the three components, an instrument development was performed based on the recommendation by Wilson (2005, 2008, 2009), which covers four steps: developing construct variables of students' ability focusing on one characteristic assessed at one time; developing item design of a particular task to measure students' responses; designing result and assessment space in which all students' responses are categorized in all items related to construct variables, and; performing the Rasch model measurement. All of the four steps have been applied in instrument development with different constructs (Barbera, 2013; Hadenfeldt et al., 2016; Laliyo et al., 2022; Lu & Bi, 2016; Mulford & Robinson, 2002; Pentecost & Barbera, 2013; Wei et al., 2012; Wind et al., 2018), and now are implemented in developing the instrument of the present work.

Table 2. Construct of scientific explanations on chemical phenomena

No.	Chemical phenomena	Pre-requisite knowledge	Item	Grade
1	Rusting iron	Redox	A1	A=10
2	Fruit rot	Redox	A2	
3	Color changes in apple	Redox	A3	
4	Changes in rotting banana	Redox	A4	
5	Coal formation	Hydrocarbon	B5	B=11
6	Garbage decomposition	Hydrocarbon	B6	
7	Petroleum formation	Petroleum	B7	
8	Wood burning	Thermochemistry	B8	
9	Photosynthesis	Thermochemistry	B9	
10	Water evaporation	Thermochemistry	B10	
11	Melting ice cubes	Thermochemistry	B11	
12	The process of acid rain	Acid-base	B12	
13	Formation of CO_2 from baking soda & vinegar	Acid-base	B13	
14	Antacids for stomach ulcers	Acid-base	B14	
15	The use of detergent	Hydrolysis	B15	
16	The use of fertilizer	Hydrolysis	B16	
17	The use of bleach on cloth	Hydrolysis	B17	
18	Hydrolysis process	Hydrolysis	B18	
19	Weathering process	Hydrolysis	B19	
20	Blood pH regulation	Buffer solution	B20	
21	Deficiency in red blood cells in the body	Buffer solution	B21	
22	The process of dissolving salt in water	Solubility & solubility product	B22	
23	Drinking water purification process	Colloid	B23	
24	The use of sunscreen	Colloid	B24	
25	Electroplating on metal	Elemental chemistry	C25	C=12
26	Fireworks	Elemental chemistry	C26	
27	Fireworks flaming colors	Elemental chemistry	C27	
28	Food preservatives	Benzene and its derivatives	C28	
29	Fermentation process	Macromolecule	C29	
30	The use of shallot	Macromolecule	C30	

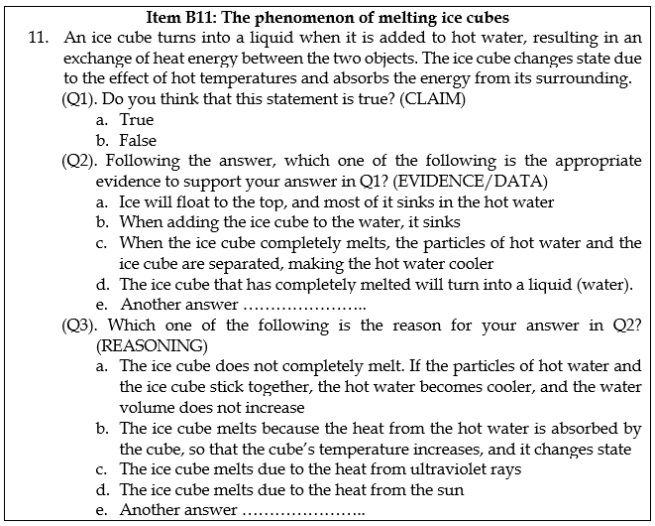
Determining construct variable: Constructing scientific explanations

The focus of this first step was to develop the construction of the assessed variable, i.e., constructing scientific explanations on chemical phenomena. A scientific explanation is a logical explanation of a phenomenon based on scientific facts and forming relationships based on evidence and logical reasoning (Berland & Reiser, 2009; Wang, 2015). A strict psychometric standard was used as the guideline in developing this construct (Kane, 2016; Van Vo & Csapó, 2021), which measures in stages the ability of students to make claims of correct knowledge/understanding (Q1), propose evidence or concepts that support their knowledge claims (Q2), and to make logical reasoning that explains the relationship between claim and evidence (Q3). The substance of each construct embodies chemical phenomena issues in which its problem-solving requires the mastery of basic chemistry concepts (e.g., acid and base). This has been stipulated in the standard of chemistry curriculum in senior high schools 2013 and the regulation of the Ministry of Education and Culture of the Republic of Indonesia number 37 of 2018. As many as 30 chemical phenomena were identified from the curriculum review results and teachers' interviews. Furthermore, these phenomena were studied by students in a formal situation in senior high schools. The results are provided in **Table 2**.

Item designing and assessment

The second step was the design of the test item and assessment. Assessments take performance-based assessments format, while the test item is in an open multiple choice-question format. Such a format is widely used in measuring scientific reasoning (Opitz et al., 2017). This is due to the effectiveness and simplicity of the test format in measuring skills (Briggs, 2009; Wilson, 2008; Zhan et al., 2017) and the lesser effect compared to other instruments (Schwichow et al., 2016). Consequently, the measurement accuracy of respondents with large numbers is higher, which culminates in better data collection and assessment (Van Vo & Csapó, 2021).

In diagnosing the framework of understanding, the two-tier multiple choice questions by Treagust (1988) are preferable to be applied in the present work. The model by Treagust (1988) was then modified into three-tier multiple choice questions. An example of the test is displayed in **Figure 1**. Item B11 measures the construction of scientific explanations of melting ice



Item B11: The phenomenon of melting ice cubes

11. An ice cube turns into a liquid when it is added to hot water, resulting in an exchange of heat energy between the two objects. The ice cube changes state due to the effect of hot temperatures and absorbs the energy from its surrounding.
(Q1). Do you think that this statement is true? (CLAIM)
   a. True
   b. False
(Q2). Following the answer, which one of the following is the appropriate evidence to support your answer in Q1? (EVIDENCE/DATA)
   a. Ice will float to the top, and most of it sinks in the hot water
   b. When adding the ice cube to the water, it sinks
   c. When the ice cube completely melts, the particles of hot water and the ice cube are separated, making the hot water cooler
   d. The ice cube that has completely melted will turn into a liquid (water).
   e. Another answer ……..
(Q3). Which one of the following is the reason for your answer in Q2? (REASONING)
   a. The ice cube does not completely melt. If the particles of hot water and the ice cube stick together, the hot water becomes cooler, and the water volume does not increase
   b. The ice cube melts because the heat from the hot water is absorbed by the cube, so that the cube's temperature increases, and it changes state
   c. The ice cube melts due to the heat from ultraviolet rays
   d. The ice cube melts due to the heat from the sun
   e. Another answer ……..

Figure 1. Item B11 sample: Phenomenon of melting ice cubes (Source: Authors' own elaboration)

cubes. Question Q1 of this item, consisting of two options, i.e., true (T) and false (F), which examines the claim of concept mastery of ice cube transformation and the melting process, due to heat absorption. Question Q2 measures evidence or data claim of Q1. There are five options: one correct answer, three distractors, and one open-answer response. Students are free to provide their own answers in open-ended response formats, and these responses can be selected and scored based on their content. Question Q3 measures students' reasoning that depicts the correlation between Q1 and Q2. This question has five options similar to Q2.

Distractors in Q2 and Q3 enhance the item diagnostic capabilities (Herrmann-Abell & DeBoer, 2011; Sadler, 1998) and minimize correct answers by guessing (Herrmann-Abell & DeBoer, 2011; Lu & Bi, 2016; Sadler, 1998). The probability of students guessing the correct answer in Q1 is 50:50, while in Q2 and Q3 is only 0.20. The distractor is an option that seems correct but is conceptually unacceptable due to its scientific contradictions or misconceptions (Herrmann-Abell & Deboer, 2016; Laliyo et al., 2020; Wind & Gale, 2015). Students' responses to each item (Q1, Q2, and Q3) are then evaluated and categorized based on the assessment rubric (**Table 3**). For example, a correct response on items Q1, Q2, and Q3 are labeled CCC. This code signifies that the student has a high ability level of constructing scientific explanations (labeled as CSEH) and scored three. If the label reads "CIC", the student has a moderate level of constructing scientific explanations (labeled as CSELR) and scored one. In this category, although the Q3 response is correct, it is unscientific due to the incorrect response of Q2.

Outcome space and data collection

The third stage is outcome space and data collection. Outcome space states the correlation between items and construct maps (Bond & Fox, 2007; Wilson, 2009). This is a content validity testing (construct) performed independently by two chemistry education experts and three senior high school chemistry teachers. These five validators were assigned to assess the correlation between the answer choices for questions (Q1, Q2, and Q3) in each item with students' abilities in constructing scientific explanations. They ensured that the questions were easy to understand, and students would not give incorrect answers only because of poor language mastery. In addition, the validators were asked to ensure that the questions (Q1, Q2, and Q3) in each item follow the measured construct and are unambiguous, and the time allocation is sufficient. Questions do not direct students to one of the answer choices and do not contain subjective or emotional words. After correcting the items according to the validators' recommendations, an analysis of the validators' approval information was carried out using the Fleiss' K measure. The value of κ=.98 was obtained in which $p<.0001$, meaning that all validators agree that test items have good validity in linking the answer choices with the students' ability in constructing scientific explanations (Landis & Koch, 1977).

All of data in this research were collected directly in the school and classroom for four months. The data collection schedule was adjusted to the school and student class schedules. Permission from the school and faculty was given prior to data collection as they are responsible for overseeing and managing the educational activities of students. Additionally, schools and faculties are legally responsible for the well-being and safety of their students while they are on school premises or participating in school-related activities. The

Table 3. Assessment rubric

Question			Code	Score	Category
Q1	Q2	Q3			
Correct	Correct	Correct	CCC	3	High level of ability in constructing scientific explanations (HACSE)
Correct	Correct	Incorrect	CCI	2	Moderate level of ability in constructing scientific explanations (MACSE)
Correct	Incorrect	Correct	CIC	1	Low level of ability in constructing scientific explanations (LACSE)
Correct	Incorrect	Incorrect	CII		
Incorrect	Correct	Correct	ICC	0	No ability in constructing scientific explanations (NACSE)
Incorrect	Correct	Incorrect	ICI		
Incorrect	Incorrect	Correct	IIC		
Incorrect	Incorrect	Incorrect	III		

process of students giving responses was carried out for 45 minutes. All students were asked to answer all questions in the measurement instrument. Most of them completed the test in less than 45 minutes. Each student was given a set of test items and a written answer sheet. After the test, all questions and answer sheets (same in number) were collected.

Measurement of Rasch partial credit model

At this stage, the relationship between the scores and the measured variables was defined applying the Rasch PCM measurement. The Rasch PCM is an extension of the dichotomous Rasch model (Rasch, 1980; Sumintono & Widhiarso, 2015) for the characteristics of polytomous data (Eggert & Bögeholz, 2010). This model assumes that partial success on certain test items is expressed by partial credit. Besides, responses given by partial credit are hierarchical, implying that responses given by higher partial credit are qualitatively better than those given by lower partial credit. This procedure is very useful for procedural competency assessments like constructing scientific explanations or problem-solving abilities in which students' answers are not only marked as correct or incorrect, but also are assessed by category. Thus, it can provide a more detailed overview of students' ability. The present study intends to reveal the difference in students' abilities in constructing scientific explanations and to diagnose the item difficulty level. The Rasch PCM was employed to describe the different categories of students' abilities continuously, from no constructing scientific explanations ability to a high level of constructing scientific explanations ability.

Data Analysis

The measurement results were still ordinal data. These data were converted to interval data with the same logit scale, using the WINSTEPS 4.5.5 software (Bond & Fox, 2015; Linacre, 2020). The result was a data calibration of the levels of student's ability and item difficulty in the same interval. Test development involved investigating the evidence for uni-dimensionality, reliability, fit statistics, and item quality testing. This is a statistical data analysis technique in evaluating the learning process and cognitive potential in educational settings (Stevenson et al., 2013). Rasch modeling was applied in this research as the primary psychometric approach in evaluating students' abilities in constructing scientific explanations of chemical phenomena. Consequently, in terms of items, the higher the logit score, the better the constructing scientific explanations abilities.

A one-way ANOVA was used to test the difference in students' abilities in different classes. Furthermore, this study also relied on the differential item functioning (DIF) analysis to examine the characteristics of items. This is evidence of measuring invariance by comparing the different abilities of members of separate groups, to test whether an item is fair or unbiased between classes. The testing result can compare students' performances based on the difference in demography to the item level.

RESULTS AND DISCUSSION

This section elaborates on the results of analysis and test validity of constructing scientific explanations abilities regarding phenomena using the Rasch model, the evaluation result of constructing scientific explanations abilities, the result of identifying test items whose functions differ in terms of sex and hometown, result of measuring the difference in item difficulty, and explaining why students' difficulty levels are different.

RQ1. To What Extent Do the Data Collected Using the Instrument Developed in This Study Fit the Rasch Model?

The estimation of data fit with Rasch modeling is based on testing uni-dimensionality, reliability, fit statistics, item fit order, and Wright map. First, uni-dimensionality is the main requirement for Rasch measurement (Bond & Fox, 2007; Chi et al., 2021; Wang & Willson, 2005). Uni-dimensionality is a measure to ensure that the developed test instrument is able to measure the construct, meaning that the item only measures one construct at a time, namely scientific argumentation skills (Bond & Fox, 2007). Uni-dimensionality measurement, in the present work, uses the principal component analysis (PCA) of residuals to estimate the extent to which instrument diversity measures what it is supposed to measure (Aryadoust et al., 2021; Ding, 2018; Sumintono & Widhiarso, 2014; Tseng et al., 2019). If the measurement result shows data that closely fit the Rasch model, most of the non-random variance (not randomized) found in the data can be explained by one latent dimension (Chi et al., 2021; Eckes, 2015). The result of the variance measurement (raw variance explained by measures) of this research data gets 27.7%, which is almost the same as the expected value of 27.4%. This indicates that the minimum uni-dimensionality requirement of 20% can be met. Moreover, the unexplained variance values obtained by the instrument are all below 7.0%, whereas the ideal value does not exceed 15.0%. This confirms that the item independence level in the instrument is good; 30 items tend to measure a single latent trait (Linacre, 2020; Ling Lee et al., 2020; Sumintono & Widhiarso, 2015).

Second is reliability. The Rasch model provides two reliability statistics: reliability index and separation index. Separation index is a measure of the relative approximation spread to accuracy of the measure (Chi et al., 2021). The measure of this separation index can reach much higher values, depending on the measure of the error variance (Eckes, 2015). These reliability statistics show reliability of measurement differences.

Table 4. Summary of fit statistics

Measures (logit)	Student (n=703)	Item (n=30)
Mean	.37	.00
SE (standard error)	.21	.04
SD (standard deviation)	.56	.43
Outfit mean square (mean)	1.02	1.02
Infit mean square (mean)	1.02	1.00
Separation index	2.25	9.68
Reliability index	.83	.99
Cronbach's alpha (KR-20)	.84	

It is seen from **Table 4** that the person reliability index reaches logit +.83, and item reliability index arrives at logit .99. Both indexes are categorized as good as they get higher values than logit .8 as recommended by Bond and Fox (2007). On the other hand, the person separation index and item separation index obtain logit 2.25 and logit 9.68, respectively. Both indexes have met the recommended criterion of logit 2.0 Linacre (2020). The person separation index of logit 2.25 and person reliability index of logit .83 reflect the adequate sensitivity of the instrument, particularly in distinguishing between students with a high and a low level of abilities. The mean of person logit of logit .37 signifies that all students have abilities above the mean of item logit of logit .00. Next, the standard deviation of person is logit .56, indicating a fairly wide level of dispersion of students' abilities. The item separation index of logit 9.68 and item reliability index of logit 9.99 become empirical evidence of students' ability levels and support an excellent instrument construct validity (Boone et al., 2014; Boone & Staver, 2020; Linacre, 2020; Sumintono & Widhiarso, 2014).

Third, fit statistics is used to ensure the validity of the test construct (Banghaei, 2008; Chan et al., 2021). Its function is to estimate whether or not the item fits the model and in accordance with the concept of a single attribute (Boone et al., 2014; Boone & Noltemeyer, 2017; Boone & Staver, 2020). Item estimation is based on the value of mean square residual (MNSQ). A value that shows how significant the impact of the fit discrepancy is, with two forms: outfit MNSQ and infit MNSQ. Outfit is a Chi-square that is sensitive to outliers. Outliers are answers that happen to be correct (guessing) of low-ability students or incorrect answers due to the carelessness of high-ability students. Next is the value of infit MNSQ. This value is affected by response patterns close to the item difficulty or students' ability. The ideal value of infit and outfit MNSQ is 1.0, ranging from 0 to infinity (Eckes, 2015). MNSQ value of more than 1.0 indicates that the item has unexpected variability (misfit or underfit). In contrast, MNSQ value of less than 1.0 implies that the item is easy to predict (overfit). Underfit is generally considered more problematic than the overfit (Myford & Wolfe, 2004). This research follows the recommendation from Bond and Fox (2007) in which MNSQ value ranging from 0.7 to 1.3 reflects a good fit between the item and the model.

The measurement result shows that the value of infit and outfit MNSQ (**Table 4**) ranges from 1.00 to 1.02, meaning that the items fit the model. Hence, the items of this instrument are productive for measurement and have a logical prediction. This is strengthened by the test reliability value of raw scores of Cronbach's alpha (KR-20) with logit .84. This value signifies that 703 students and 30 items have good interaction between each other. Simply put, psychometric internal consistency of this research instrument is excellent and reliable (Adams & Wieman, 2011; Boone & Staver, 2020; Sumintono & Widhiarso, 2015).

Fourth is the item fit order. Item fit order testing is employed to estimate the item quality (Boone & Staver, 2020; Linacre, 2020), while ensuring the quality of the process of student responses to items (Lewis, 2022). Item is considered misfit it it does not meet the following criteria: outfit mean square residual (MNSQ): $.5<y<1.5$; outfit standardized mean square residual (ZSTD): $-2<Z<+2$; and point measure correlation (PTMEA CORR): $.4<x<.8$. The value of PTMEA CORR shows the correlation between item score and person measure. The value is used to check whether or not all items are functioning as expected. The value must be positive and not close to zero (Bond & Fox, 2015; Boone & Staver, 2020). If a positive value is obtained, the item is considered acceptable. On the contrary, if a negative value is obtained, the item is not functioning well or comprises misconception (Bond & Fox, 2015; Boone et al., 2014; Sumintono & Widhiarso, 2015). The result of item fit measurement (**Table 5**) reveals that all items have met the criteria mentioned earlier, and there is no negative value for PTMEA CORR. This implies that all

Table 5. Item fit analysis

Item	Measure	Infit	Outfit		PTMEA
		MNSQ	MNSQ	ZSTD	CORR
A2	-.40	1.43	1.46	9.17	.37
A1	-.27	1.23	1.24	5.27	.35
B15	-.96	1.22	1.21	3.04	.40
A3	-.11	1.17	1.20	4.49	.39
B23	-.72	1.17	1.14	2.58	.50
B10	.28	1.11	1.16	3.18	.44
C26	-.48	1.16	1.16	3.35	.50
B17	-.57	1.14	1.12	2.36	.45
B16	.33	1.07	1.09	1.74	.33
B13	.14	1.04	1.07	1.54	.45
C25	.55	.97	1.07	1.38	.30
B5	.45	.91	1.06	1.16	.14
B19	.29	1.02	1.06	1.20	.43
B24	-.40	1.05	1.05	1.18	.57
B12	-.72	1.01	.99	-.15	.37
B9	.36	.91	1.00	-.01	.28
B7	.46	.96	.99	-.22	.35
B22	.21	.95	.99	-.16	.42
B6	.08	.97	.98	-.35	.54

Table 5 (Continued). Item fit analysis

Item	Measure	Infit	Outfit		PTMEA
		MNSQ	MNSQ	ZSTD	CORR
B11	.42	.92	.98	-.75	.25
B14	.16	.95	.96	-.80	.41
B21	-.22	.89	.92	-1.88	.51
B8	.18	.89	.91	-2.01	.36
B20	.45	.89	.91	-1.72	.42
C28	-.45	.91	.88	-2.88	.58
C30	.08	.88	.88	-2.84	.55
A4	-.33	.85	.87	-3.12	.53
C29	.23	.83	.86	-3.24	.54
C27	.26	.80	.83	-3.92	.44
B18	.71	.56	.61	-8.00	.38

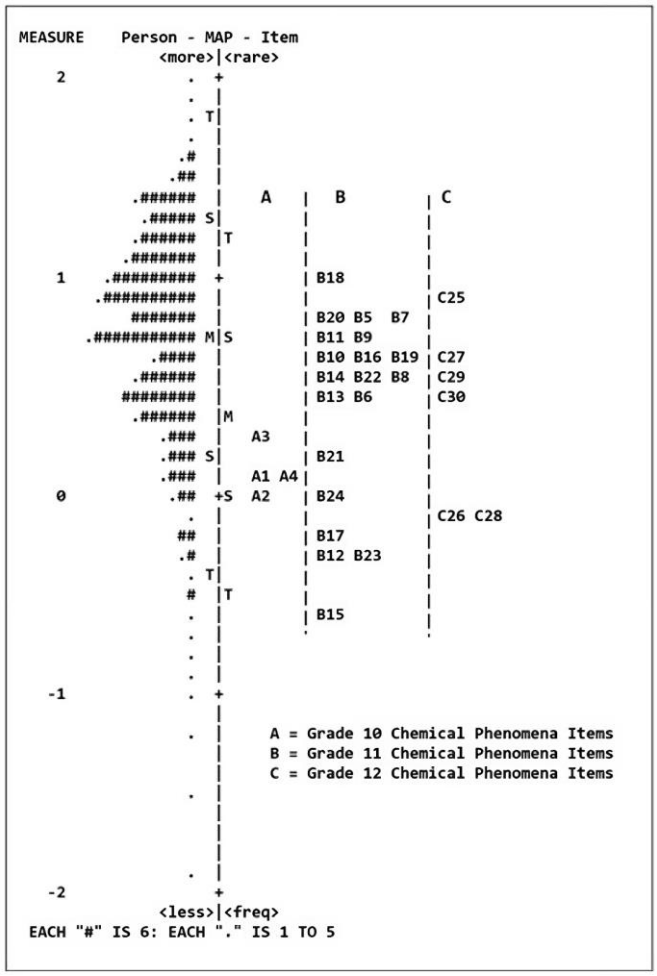


Figure 2. Wright map (Source: Authors' own elaboration)

items are functioning well and meet the qualifications of good item quality (Bond & Fox, 2015; Boone & Staver, 2020).

Fifth is the Wright map (**Figure 2**) displaying a graphical representation of the distribution of students' ability (left side) and item difficulty level (right side), which are on the same logit scale (Bond & Fox, 2015). This map measures the consistency between the levels of students' ability and item difficulty. The higher the logit scale, the higher the levels of students' ability and item difficulty. Conversely, the lower the logit scale, the lower the levels of students' ability and item difficulty (Boone et al., 2014). The logit size for each item reflects the students' ability to construct scientific explanations related to the measured construct. Indirectly, it can be interpreted that the acquisition of a small logit item size reflects the students' inability to understand the measured phenomena. Thus, it can be assumed that they tend to have limited knowledge related to the conceptual knowledge learned in formal classes.

Wright map has illustrated that all items in the measurement instrument cover most of the students' ability and have the same difficulty level. However, in students' ability lower than logit -.96 and higher than logit .71, there is no item equivalent to the ability level in question, which needs further investigation. Most items tend to be in the middle of the map. Some items have the same logit. There are also items with the same difficulty level, such as item B20, item B5, item B10, item B16, item B19, and item C27. Item B18 (.71) is the item with the highest difficulty level, whereas item B15 (-.96) is the one with the lowest difficulty level. The series of results of the validity and reliability tests that have been carried out indicate that the data collected from the measurement instruments developed in this study are in accordance with the Rasch modeling.

RQ2. How Do the Students' Ability to Construct Scientific Explanations on Chemical Phenomena, With Regard to Different Classes, Sex, and Hometown?

Figure 3 presents the visualization of the difference in students' scientific argumentation skills from different classes. It is clear that students in class J have better average scores (.97) than those in class I (.77), A (.60), H (.54), C (.50), E (.37), F (.06), D (.05), and B (-.10). The ability of students in classes H, I, and J is better possibly because they benefit from a longer learning experience. Moreover, the difference between classes is tested by relying on one-way ANOVA test. The test determines the difference in ability, especially for data that are presumed to break the assumptions of normality and homogeneity (Chi et al., 2018; Liu & Boone, 2006). The testing result suggests that there is a difference in the ability of students in classes ABCDEFGHIJ, with a significance value <0.05 (Sig.=0.000). Further testing using Bonferroni's post hoc test finds out that statistically, students' ability in classes A, C, E, G, H, I, and J is significantly different; yet the ability of students in classes B, D, and F is insignificantly different.

At the level of individual (students), the testing of the difference in scientific argumentation skills is performed by using the measurement result of logit value of person (LVP). Following the person mean (.37) and standard deviation (.56), LVP of all students is grouped into four, as follows:

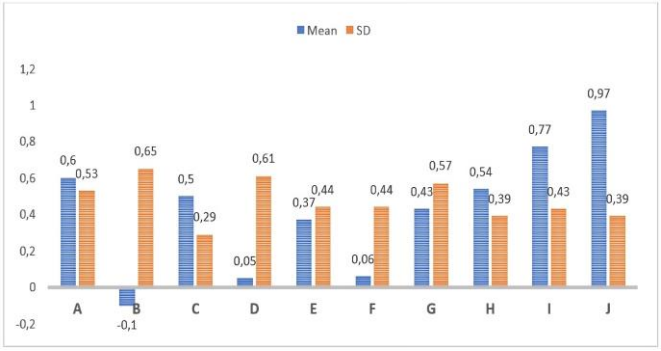


Figure 3. Average of students' ability in classes A to J (Source: Authors' own elaboration)

(a) a group of students with a very high level of ability (LVP≥.93),

(b) a group of students with a high level of ability (.37≥LVP≥.93),

(c) a group of students with a moderate level of ability (-.19>LVP>.37), and

(d) a group of students with a low level of ability (LVP≤-.19).

The highest percentage is achieved by the group of students with a very high level of ability; those in class A (3.41%), class J (3.12%), and class I (2.98%). In the group of students with a high level of ability, they are dominantly in class A (14.20%), class G (5.38%), and class F (4.83%). Next, in the group of students with a moderate level of ability, they are mostly in class F (9.53%), class B (4.55%), and class A (4.41%). Lastly, in the group of students with a low level of ability, they are mostly in class B (4.41%), class F (4.13%), and class A (2.13%). In terms of sex, female students are better than male ones. The majority of female students have very high (10.2%), high (37.4%), moderate (22.2%), and low (7.58%) levels of abilities. In the matter of students with a very high level of ability, students who come from Bone Bolango (7.54%) are better than those from Gorontalo (4.13%) and Limboto (1.13%). Overall, more than half of the students are in groups of a high level of ability (58.60%), followed by a moderate level of ability (29.00%), a low level of ability (13.80%), and a very high level of ability (12.80%). Such facts signify that the distribution of students' ability is relatively different and uneven in each class, and no class has the ability in certain LVP categories (**Table 6**).

Table 6. Logit value of person (LVP) analysis (n=703)

Demography	Logit value of person (%)			
	VH (LVP≥ .93)	High (.37≥LVP ≥.93)	M (-.19≥ LVP≥.37)	Low (LVP≤ -.19)
Students				
Class A	24 (3.41)	100 (14.2)	31 (4.41)	15 (2.13)
Class B	5 (0.71)	15 (2.13)	32 (4.55)	31 (4.41)
Class C	0 (0.00)	10 (1.42)	2 (0.28)	2 (0.28)
Class D	0 (0.00)	4 (0.56)	15 (2.13)	3 (0.43)
Class E	2 (0.28)	28 (3.98)	17 (2.41)	6 (0.85)
Class F	6 (0.85)	34 (4.83)	67 (9.53)	29 (4.13)
Class G	4 (0.56)	41 (5.83)	18 (2.56)	10 (1.42)
Class H	6 (0.85)	24 (3.41)	7 (0.99)	1 (0.14)
Class I	21 (2.98)	28 (3.98)	8 (1.13)	1 (0.14)
Class J	22 (3.12)	28 (3.98)	7 (0.99)	0 (0.00)
Total	90 (12.8)	412 (58.6)	204 (29.0)	97 (13.8)
Sex				
Male	18 (2.56)	49 (6.97)	48 (6.82)	43 (6.1)
Female	72 (10.2)	263 (37.4)	156 (22.2)	54 (7.58)
Hometown				
Gorontalo	29 (4.13)	129 (18.3)	80 (11.4)	51 (7.25)
Limboto	8 (1.13)	62 (8.82)	84 (11.9)	34 (4.83)
Bone Bolango	53 (7.54)	121 (17.2)	40 (5.69)	12 (1.70)

Note. VH: Very high & M: Moderate

RQ3. Based on the Measurement Result in the Item Level, Are There Any Items That Have Different Functions With Regard to Sex and Hometown of the Students?

Testing items whose functions differ among students of different sex and hometown relied on DIF analysis. An item is considered to have DIF if fulfilling three criteria:

(a) having the value of t of less than -2.0 or greater than 2.0,

(b) having the DIF contrast value of less than -.5 or greater than .5, and

(c) having the p (probability) value of less than -.05 or greater than .05 (Bond & Fox, 2015; Boone et al., 2014; Chan et al., 2021).

The result of testing DIF items based on the difference in sex and hometown of the students does not find items that meet the three criteria mentioned previously. In other words, no items are responded differently by male and female students from different hometowns; all items have the same degree of difficulty.

RQ4. From the Measure of Item Logit, in What Ways Does the Students' Difficulty Level Differ From Each Other in Constructing Scientific Explanations on Chemical Phenomena?

Table 7 presents the grouping of the logit value of item (LVI). The grouping process is based on the mean of person (.00) and standard deviation (.43), divided into four groups of item difficulty levels, namely:

(a) a group of the most difficult items (LVI≥.43),

(b) a group of difficult items (.00≥LVI≥.43),

(c) a group of easy items (-.43≥LVI≥.00), and

(d) a group of the easiest items (LVI≤ logit .-43).

Table 7 shows some interesting cases regarding the distribution of item difficulty level based on students' different classes. First, the item distribution in each class tends to be different; there is no item considered difficult

Table 7. Logit value of item (LVI) analysis (n=30)

Class (mean)	Item			
	Very difficult: LVI≥.43	Difficult: .00≥LVI≥.43	Easy: -.43≥LVI≥.00	Very easy: LVI≤-.43
A (.60)	B5, B7, B8, B9, B18, B20, C25	B10, **B11**, B16, B19, B22, C27, C29	A1, A3, B6, B13, B14, B21, C30	A2, A4, B12, B15, B17, B23, B24, C26, C28,
B (-.10)	B9, B14, B18, B22, C27, C29	B7, B10, **B11**, B19, B20, B21, B25, C30	A2, A3, A4, B5, B6, B8, B13, B16, B24, C28	A1, B12, B15, B17, B23, C26
C (.50)	A1, A4, B5, B8, B18, B20, B21, B25	**B11**, B13, B16, B17, B22, C29	A2, B7, B9, B10, B12, B14, B19, C27, C30	A3, B6, B15, B23, B24, C26, C28
D (.05)	B7, B9, B14, B16, B18, B19, B20	B5, B8, B10, **B11**, B22, B25, C27, C30	A1, B6, B12, B13, B21, B24, C26, C28, C29	A2, A3, A4, B15, B17, B23
E (.37)	B5, B9, B10, **B11**, B18, B25	B7, B14, B16, B19, B20, B22, B25, C29, C30	A1, A2, A3, A4, B6, B8, B12, B13, B21, B24, C27, C28	B15, B17, B23, B24, C26
F (.06)	B10, B13, B16, B19, B20, B25, C29	A3, B5, B7, B8, **B11**, B14, B18, B22, C27, C30	A2, A4, B6, B9, B21, C28	A1, B12, B15, B17, B23, C26
G (.43)	B5, B7, B13, B14, B18, B19, B20, B25	B8, B9, B10, **B11**, B16, C27, C30	A3, A4, B6, B17, B21, B22, B23, C28, C29	A1, A2, B12, B15, B24, C26
H (.54)	B10, **B11**, B18, B25	B5, B7, B8, B13, B16, B19, B20, B21, B22, C27, C29, C30	A1, A2, A3, B6, B9, B14, B23, , C28, C30	A4, B12, B15, B17, B24, C26
I (.77)	B5, B7, **B11**, B16, B18, B25	A1, B8, B9, B10, B14, B19, B20, B22, C27	A2, A3, A4, B6, B13, B24, C26, C29, C30	B12, B15, B17, B21, B23, C28
J (.97)	B7, **B11**, B16, B18, C27	A3, B5, B8, B13, B14, B19, B25	A1, A2, A4, B9, B10, B17, B20, B21, B22, C26, C29, C30	B6, B12, B15, B23, B24, C28

by all classes. Item B11 (.42) is considered to have the highest difficulty level by classes E, H, I, and J; meanwhile, to classes A, B, C, D, F, and G, this item is considered difficult.

Another example is item B18 (.71). This item is considered the most challenging in classes A, B, C, D, E, F, G, H, I, and J.

Second, if it is based on the average score of students' ability by class, certain items should have been considered easy. Nevertheless, in some classes, the items are categorized as the most difficult ones, e.g., item B11 (.42). The measure of this item is less than the average score of students' ability in class J (.97), class I (.77), and class H (.54), meaning that the logit of the class average ability greater than the measure of item B11 should have enabled the item to be categorized easy in class J, I, and H. In fact, this item is the most difficult one, although the university students in class J, I, and H are from the chemistry department with more learning experiences than those in high school classes, specifically in learning the phenomenon of melting ice cubes. In addition, it was found that based on the average score, first-year university chemistry students (class H) tended to have the same abilities as classes A, C, E, and G. In contrast, the ability of classes B, D, and F students was significantly lower. These two examples of facts show the level of difficulty of items in each class is different. Even though university students (HIJ class) have a longer learning experience, it does not mean that they can construct better scientific explanations than students. Allegedly, this tends to be determined by how students construct their ideas. According to Aktan (2013), Emden et al. (2018), and Hadenfeldt et al. (2016), each student tends to have a different way of building scientific understanding and explanation. Differences in ways of understanding scientific phenomena cause the construction of students' understanding to be often different, at the same time changing according to the understanding they form themselves, and often varying at different levels of understanding. This notion underlines the causes of the different levels of understanding of each student as understanding is obtained and formed in a way that is not the same or not linear (Hadenfeldt et al., 2013; Neumann et al., 2013). This fact reinforces the research conclusions reported by Yao and Guo (2018) that the abilities of students in constructing scientific explanations can be different, where these differences reflect the conditions and learning practices held in each school.

RQ5. Based on the Item Response Patterns, What Causes the Different Difficulty Levels of Students in Constructing Scientific Explanations About Chemical Phenomena?

Responding to why students have different item difficulty levels can be analyzed using the response patterns in each item (Please see **Appendix A** and **Appendix B**). **Figure 4** provides the response patterns of ten students with a high level of ability. Student 115 (1.15), 283 (.89), and 355 (1.24) give responses of three for item B11 (.42). According to **Table 3**, response 3 is the response pattern of students with a high level of constructing scientific explanations ability, indicating that these three students give correct responses

GUTTMAN SCALOGRAM OF RESPONSES:

Person	Item							
	1121222 2 3 11 222111	1	2 21			Person	Mean	
	52376842411306348297096 9	1	05758	Responden ID	Class	Measure	Class	
		-	-----					
2	+333333333123133311313131	1	23113	002WPAAATSTST	A	1.15	.60	
115	+221313333131331311213311	3	21311	115APAAASSTST	B	1.15	-.10	
263	+233133331113232111333233	1	11311	263FPAACSSTST	C	.89	.50	
283	+333330333333113133331201	3	10111	283ALAADTSTST	D	.89	.05	
308	+211333333333331123331131	1	31111	308RPABESTTSS	E	.98	.37	
355	+323333333333321123133101	3	33311	355APABFSSTST	F	1.24	.06	
482	+333333333333333333323113	1	13111	482MPACGSSTTT	G	1.50	.43	
578	+323333331303233311333333	1	33311	578SPBCHSSTST	H	1.50	.54	
622	+323333331333333323332333	1	31132	622IPBCISSYST	I	1.69	.77	
699	+323333333333333121233133	1	31332	699ZLBCJSSYST	J	1.56	.97	

Figure 4. Guttman scalogram of responses (Source: Authors' own elaboration)

regarding the questions of claim (Q1), evidence (Q2), and justification (Q3), labelled as CCC. On the other hand, student 2 (1.15), 263 (.89), 308 (.98), 482 (1.50), 578 (1.50), 622 (1.69), and 699(1.56) provide a response of 1. Response 1 is the response pattern of students with a low level of constructing scientific explanations ability, implying that these seven students respond correctly in Q1, yet incorrectly in Q2 and Q3. Thus, they cannot provide evidence and justification for their knowledge claims related to the phenomenon of melting ice.

Referring to **Table 3**, we can investigate further why students cannot show evidence or respond incorrectly in Q2. There are two possible response patterns: CIC and CII. CIC is the response pattern of correct Q1, incorrect Q2, and correct Q3. Although the Q3 is correct, it is unacceptable since the Q2 is incorrect, implying that it is impossible to be able to explain the relationship between claim and evidence correctly, yet having incorrect evidence. CII is the response pattern of correct Q1, incorrect Q2, and incorrect Q3. Accordingly, in item B11 and question Q2, the seven students choose answer:

(a) ice will float to the top, and most of it sinks in hot water,

(b) when adding ice cube to the water, it sinks, and

(c) when the ice cube completely melts, the particles of hot water and the ice cube are separated, making the hot water cooler.

These three answer choices are distractors and have misconceptions, particularly the concept of thermochemistry. This fact strengthens the assumption that the seven students do not master the basic concepts of thermochemistry competently.

Why do students, although having a high-level ability, have difficulty in constructing scientific explanations about chemical phenomena (itemB11)? They should be able to answer correctly because the logit size of their ability is greater than the size of item B11 (.42 logit). This case can be explained as follows. First, the emergence of difficulties can be caused by gaps in understanding (Kapici & Akcay, 2016) as a result of the existence of an alternative conceptual framework formed by students in a way that they can understand themselves (Lu & Bi, 2016; Yildirir & Demirkol, 2018). These alternative frameworks usually contain misconceptions (Johnstone, 2006, 2010; Taber, 2002, 2009) and even tend to be permanent (Hoe & Subramaniam, 2016; Laliyo et al., 2022). Several previous studies have concluded that students tend to experience misconceptions (Alamine & Etokeren, 2018; Yasar et al., 2014) even though they have experienced formal learning (Allen, 2014; Suharto & Csapó, 2021). Students' efforts to understand the phenomenon of melting ice cannot be separated from how their efforts in constructing a correct understanding of the concept of abstract particle properties (Johnstone, 1991), which do not look real (Stojanovska et al., 2012), making it difficult for them to understand by means of the usual (Cheng, 2018; Johnstone, 2010).

Second, it is possible to derive from the relatively weak conditions and practice of learning instructions in forming a meaningful understanding, which explains the relationship between macroscopic, submicroscopic, and symbolic representations (Chi et al., 2018; Chittleborough et al., 2005; Davidowitz et al., 2010; Gilbert & Treagust, 2009). It means that the epistemological experience of students in learning tends to be weak. Less-meaningful learning can also originate from the limited ability of teachers to carry out learning instructions based on the discovery of claims, evidence, and justification (Yao & Guo, 2018). In addition, less-meaningful understanding of students tends to be caused by how they store information. Learners often do not store information that is not used and tend to need relatively sufficient time to refresh their previously formed knowledge and skills (Clark et al., 2022).

A number of the results of this study indicate that the involvement of students in the learning process and scientific practice has distinctive, epistemic characteristics (Deng & Wang, 2017; Driver et al., 2000; Erduran et al., 2004; Osborne et al., 2004). This is probably related to the students' ability with their epistemological experiences in learning. A study reported by Dillon et al. (2006) and Rickinson et al. (2004) describes that although the learning process is performed in laboratory practice and in the field, if students are not allowed to make more profound meaning of learning and think critically, they cannot make cognitive and/or affective connections with what they are learning. Research by Jin et al. (2021) showed that the level of meaningfulness of students' scientific explanations or arguments can be empirically proven to be different and is relatively determined by how involved they are in learning. Thus, this is all about meaningful involvement in which students are engaged epistemologically in contextualizing and forming meaning for their learning. Such a meaning-making aspect of the learning experience is referred to as epistemic engagement (Ryder & Leach, 1999).

The presence of students with a high level of ability, yet fail in constructing scientific explanations on chemical phenomena, is possibly because there is a

misconception and the aspect of epistemic engagement in the learning process is not carried out properly. Cooper (2012) and Kinslow et al. (2018) argue that students' failure to form meaning in their learning is partly due to an epistemically less interesting learning experience in which students are involved in the data collection process, yet they do not think about the function and meaning of the information collected. On this ground, it is essential to provide learning facilities to integrate complex contextual components, such as intriguing issues and themes, into students' learning experience and learning process to improve epistemic engagement and scientific literacy (Gulacar et al., 2020; Kinslow et al., 2018; Owens et al., 2019).

CONCLUSIONS

The results of this study highlight the significance of using a Rasch measurement model to assess students' abilities in constructing scientific explanations in chemical phenomena. The findings demonstrate that the test items used in this study are valid, reliable, and predictive, making them a valuable measurement instrument for evaluating students' scientific reasoning skills. Moreover, the study reveals significant differences in students' abilities across different classes, indicating the variability in mastering basic chemistry concepts.

These findings have important implications for the field of education in Indonesia and beyond. By providing evaluative information on the application of the 2013 chemistry curriculum and students' scientific argumentation skills, this research can inform the decisions of teachers, researchers, and policymakers in developing strategies to enhance students' abilities in constructing scientific explanations. The Rasch modeling-based psychometric analysis technique used in this study can serve as a valuable approach for assessing students' scientific reasoning skills in other contexts as well. Overall, this research contributes to the advancement of science education and has the potential to improve the quality of students' learning experiences in chemistry and related fields worldwide.

Research Limitations

The present study has to be seen in light of some limitations, including the number of research participants that have not reached the population in other regions in Indonesia. It is recommended for further studies to involve more students in different education stages to get a better overview of students' progress in constructing scientific explanations on chemical phenomena. Those studies are also recommended to measure students' different backgrounds and demography, such as ethnicity, basic skills and literacy of chemistry concepts. Ethnicity closely relates to learning culture and motivation, and the basic skills of chemistry associate with the achievement of scientific literacy concepts mastery. The results are shown by testing items with different functions and followed by examination through an in-depth interview to determine the reason and the situation causing them.

Author contributions: LARF, RU, RH, MKU, & MRK: project administration, resources, & funding acquisition; **LARF & RU:** conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft preparation, & writing—review & editing; & **RH, MKU, MRK, & CP:** validation & supervision. All authors have agreed with the results and conclusions.

Funding: This study was funded by the Directorate of Research and Community Service & the Directorate General of Higher Education, Ministry of Education, Culture, Research, & Technology of the Republic of Indonesia.

Acknowledgments: The authors would like to thank the Directorate of Research and Community Service & the Directorate General of Higher Education, Ministry of Education, Culture, Research, & Technology of the Republic of Indonesia for their financial support of our research through the Institute for Research and Community Service of Universitas Negeri Gorontalo, 2022.

Ethical statement: Authors stated that ethical approval for the research was obtained from the Badan Kesatuan Bangsa dan Politik Provinsi Gorontalo committee, with approval granted on 28 February 2022 (document number: 070/KesbangPol/401/II/2022). Participants' consent was obtained in accordance with the regulations of the Institutional Review Board (IRB), ensuring confidentiality of students' identities and using the information solely for scientific development purposes.

Declaration of interest: No conflict of interest is declared by authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Adadan, E., & Savasci, F. (2012). An analysis of 16-17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument. *International Journal of Science Education, 34*(4), 513-544. https://doi.org/10.1080/09500693.2011.636084

Adams, W. K., & Wieman, C. E. (2011). Development and validation of instruments to measure learning of expert-like thinking. *International Journal of Science Education, 33*(9), 1289-1312. https://doi.org/10.1080/09500693.2010.512369

Aktan, D. C. (2013). Investigation of students' intermediate conceptual understanding levels: The case of direct current electricity concepts. *European Journal of Physics, 34*(1), 33-43. https://doi.org/10.1088/0143-0807/34/1/33

Alamina, J. I., & Etokeren, I. S. (2018). Effectiveness of imagination stretch teaching strategy in correcting misconceptions of students about particulate nature of matter. *Journal of Education, Society and Behavioral Science, 27*(1), 1-11. https://doi.org/10.9734/jesbs/2018/43063

Allen, M. (2014). *Misconceptions in primary science*. Open University Press.

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6-40. https://doi.org/10.1177/0265532220927487

Bailey, C. P., Minderhout, V., & Loertscher, J. (2012). Learning transferable skills in large lecture halls: Implementing a POGIL approach in biochemistry. *Biochemistry and Molecular Biology Education, 40*(1), 1-7. https://doi.org/10.1002/bmb.20556

Banghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transaction, 22*(1), 1145-1162.

Barbera, J. (2013). A psychometric analysis of the chemical concepts inventory. *Journal of Chemical Education, 90*(5), 546-553. https://doi.org/10.1021/ed3004353

Berland, L. K., & Reiser, B. J. (2009). Making sense of argumentation and explanation. *Science Education, 93*(1), 26-55. https://doi.org/10.1002/sce.20286

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge Taylor & Francis Group.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge Taylor & Francis Group. https://doi.org/10.1088/1751-8113/44/8/085201

Boone, W. J., & Noltemeyer, A. (2017). Rasch analysis: A primer for school psychology researchers and practitioners. *Cogent Education, 4*(1). https://doi.org/10.1080/2331186X.2017.1416898

Boone, W. J., & Staver, J. R. (2020). Advances in Rasch analyses in the human sciences. In D. M. Garner (Ed.), *Advances in Rasch analyses in the human sciences* (pp. 317-334). Springer. https://doi.org/10.1007/978-3-030-43420-5_21

Boone, W. J., Yale, M. S., & Staver, J. R. (2014). *Rasch analysis in the human sciences*. Springer. https://doi.org/10.1007/978-94-007-6857-4

Briggs, D. (2009). *The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression* [Paper presentation]. The Learning Progressions in Science Conference.

Cetin, P. S. (2014). Explicit argumentation instruction to facilitate conceptual understanding and argumentation skills. *Research in Science and Technological Education, 32*(1), 1-20. https://doi.org/10.1080/02635143.2013.850071

Chan, S. W., Looi, C. K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education, 8*(2), 213-236. https://doi.org/10.1007/s40692-020-00177-2

Chandrasegaran, A. L., Treagust, D. F., & Mocerino, M. (2007). The development of a two-tier multiple-choice diagnostic instrument for evaluating secondary school students' ability to describe and explain chemical reactions using multiple levels of representation. *Chemistry Education Research and Practice, 8*(3), 293-307. https://doi.org/10.1039/B7RP90006F

Cheng, M. M. W. (2018). Students' vizualisation of chemical reactions-insights into the particle model and the atomic model. *Chemistry Education Research and Practice, 19*(1), 227-239. https://doi.org/10.1039/c6rp00235h

Cheung, D. (2008). Developing a scale to measure students' attitudes toward chemistry lessons. *Chemistry Education Research and Practice, 9*(1), 50-59. https://doi.org/10.1080/09500690802189799

Chi, S., Liu, X., & Wang, Z. (2021). Comparing student science performance between hands-on and traditional item types: A many-facet Rasch analysis. *Studies in Educational Evaluation, 70*, 100998. https://doi.org/10.1016/j.stueduc.2021.100998

Chi, S., Wang, Z., & Liu, X. (2022). Assessment of context-based chemistry problem-solving skills: Test design and results from ninth-grade students. *Research in Science Education, 53*, 295-318. https://doi.org/10.1007/s11165-022-10056-8

Chi, S., Wang, Z., Luo, M., Yang, Y., & Huang, M. (2018a). Student progression on chemical symbol representation abilities at different grade levels (grades 10-12) across gender. *Chemistry Education Research and Practice, 19*(4), 1111-1124. https://doi.org/10.1039/c8rp00010g

Chi, S., Wang, Z., Luo, M., Yang, Y., & Huang, M. (2018b). Student progression on chemical symbol representation abilities at different grade levels (grades 10-12) across gender. *Chemistry Education Research and Practice, 19*(4), 1055-1064. https://doi.org/10.1039/c8rp00010g

Chin, C., & Brown, D. E. (2000). Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching, 37*(2), 109-138. https://doi.org/10.1002/(SICI)1098-2736(200002)37:2<109::AID-TEA3>3.0.CO;2-7

Chittleborough, G. D., Treagust, D. F., Mamiala, T. L., & Mocerino, M. (2005). Students' perceptions of the role of models in the process of science and in the process of learning. *Research in Science and Technological Education, 23*(2), 195-212. https://doi.org/10.1080/02635140500266484

Clark, T. M., Dickson-Karn, N. M., & Anderson, E. (2022). Calculating the pH of a strong acid or a strong base before and after instruction in general and analytical chemistry. *Journal of Chemical Education, 99*(4), 1587-1595. https://doi.org/10.1021/acs.jchemed.1c00819

Cooper, C. B. (2012). Links and distinctions among citizenship, science, and citizen science. *Democracy and Education, 20*(2), 1-4.

Davidowitz, B., Chittleborough, G., & Murray, E. (2010). Student-generated sub-micro diagrams: A useful tool for teaching and learning chemical equations and stoichiometry. *Chemistry Education Research and Practice, 11*(3), 154-164. https://doi.org/10.1039/c005464j

Deng, Y., & Wang, H. (2017). Research on evaluation of Chinese students' competence in written scientific argumentation in the context of chemistry. *Chemistry Education Research and Practice, 18*(1), 127-150. https://doi.org/10.1039/c6rp00076b

Dillon, J., Rickinson, M., Teamey, K., Morris, M., Young, D. S., & Benefield, P. (2006). The value of outdoor learning: Evidence from research in the UK and Elsewhere. *School Science Review, 87*, 107-111.

Ding, L. (2018). Progression trend of scientific reasoning from elementary school to university: A large-scale cross-grade survey among Chinese students. *International Journal of Science and Mathematics Education, 16*(8), 1479-1498. https://doi.org/10.1007/s10763-017-9844-0

Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287-312. https://doi.org/10.1002/(sici)1098-237x(200005)84:3<287::aid-sce1>3.0.co;2-a

Duran, M., & Dokme, I. (2016). The effect of the inquiry-based learning approach on students' critical thinking skills. *EURASIA Journal of Mathematics, Science, and Technology Education, 12*(12), 2887-2908. https://doi.org/10.12973/eurasia.2016.02311a

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Eggert, S., & Bögeholz, S. (2010). Students' use of decision-making strategies with regard to socio-scientific issues: An application of the Rasch partial credit model. *Science Education, 94*(2), 230-258. https://doi.org/10.1002/sce.20358

Emden, M., Weber, K., & Sumfleth, E. (2018). Evaluating a learning progression on "transformation of matter" on the lower secondary level. *Chemistry Education Research and Practice, 19*(4), 1096-1116. https://doi.org/10.1039/c8rp00137e

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education, 88*(6), 915-933. https://doi.org/10.1002/sce.20012

Gilbert, J K & Treagust, D. (Ed.). (2009). *Multiple representations in chemical education*. Springer. https://doi.org/10.1007/978-1-4020-8872-8

Grooms, J. (2020). A comparison of argument quality and students' conceptions of data and evidence for undergraduates experiencing two types of laboratory instruction. *Journal of Chemical Education, 97*(5), 1210-1222. https://doi.org/10.1021/acs.jchemed.0c00026

Gulacar, O., Zowada, C., Burke, S., Nabavizadeh, A., Bernardo, A., & Eilks, I. (2020). Integration of a sustainability-oriented socio-scientific issue into the general chemistry curriculum: Examining the effects on student motivation and self-efficacy. *Sustainable Chemistry and Pharmacy, 15*, 100232. https://doi.org/10.1016/j.scp.2020.100232

Hadenfeldt, Jan C, Bernholt, S., Liu, X., Neumann, K., & Parchmann, I. (2013). Using ordered multiple-choice items to assess students' understanding of the structure and composition of matter. *Journal of Chemical Education, 90*(12), 1602-1608. https://doi.org/10.1021/ed3006192

Hadenfeldt, Jan Christoph, Neumann, K., Bernholt, S., Liu, X., & Parchmann, I. (2016). Students' progression in understanding the matter concept. *Journal of Research in Science Teaching, 53*(5), 683-708. https://doi.org/10.1002/tea.21312

Herrmann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. *Chemistry Education Research and Practice, 12*(2), 184-192. https://doi.org/10.1039/c1rp90023d

Herrmann-Abell, C. F., & Deboer, G. E. (2016). *Using Rasch modeling and option probability curves to diagnose students' misconceptions* [Paper presentation]. The 2016 American Educational Research Association Annual Meeting.

Hoe, K. Y., & Subramaniam, R. (2016). On the prevalence of alternative conceptions on acid-base chemistry among secondary students: Insights from cognitive and confidence measures. *Chemistry Education Research and Practice, 17*(2), 263-282. https://doi.org/10.1039/c5rp00146c

Hong, Z. R., Lin, H. shyang, Wang, H. H., Chen, H. T., & Yang, K. K. (2013). Promoting and scaffolding elementary school students' attitudes toward science and argumentation through a science and society intervention. *International Journal of Science Education, 35*(10), 1625-1648. https://doi.org/10.1080/09500693.2012.734935

Jin, H., Yan, D., Mehl, C. E., Llort, K., & Cui, W. (2021). An empirically grounded framework that evaluates argument quality in scientific and social contexts. *International Journal of Science and Mathematics Education, 19*(4), 681-700. https://doi.org/10.1007/s10763-020-10075-9

Johnstone, A. H. (1991). Why is science difficult to learn? Things are seldom what they seem. *Journal of*

Computer Assisted Learning, 7, 75-83. https://doi.org/10.1111/j.1365-2729.1991.tb00230.x

Johnstone, A. H. (2006). Chemical education research in Glasgow in perspective. *Chemical Education Research and Practice, 7*(2), 49-63. https://doi.org/10.1039/b5rp90021b

Johnstone, A. H. (2010). You can't get there from here. *Journal of Chemical Education, 87*(1), 22-29. https://doi.org/10.1021/ed800026d

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy and Practice, 23*(2), 198-211. https://doi.org/10.1080/0969594X.2015.1060192

Kapici, H. O., & Akcay, H. (2016). Particulate nature of matter misconceptions held by middle and high school students in Turkey. *European Journal of Education Studies, 2*(8), 43-58. https://doi.org/10.5281/zenodo.163547

Kinslow, A. T., Sadler, T. D., & Nguyen, H. T. (2018). Socio-scientific reasoning and environmental literacy in a field-based ecology class. *Environmental Education Research*, 4622, 1-23. https://doi.org/10.1080/13504622.2018.1442418

Laliyo, L. A. R, Sumintono, B., & Panigoro, C. (2022). Measuring changes in hydrolysis concept of students taught by inquiry model: Stacking and racking analysis techniques in Rasch model. *Heliyon, 8*, e09126. https://doi.org/10.1016/j.heliyon.2022.e09126

Laliyo, Lukman A R, Kilo, A. La, Paputungan, M., Kunusa, W. R., & Dama, L. (2022). Rasch modelling to evaluate reasoning difficulties, changes of responses, and item misconception pattern of hydrolysis. *Journal of Baltic Science Education, 21*(5), 817-835. https://doi.org/10.33225/jbse/22.21.817

Laliyo, Lukman A. R, Tangio, J. S., Sumintono, B., Jahja, M., & Panigoro, C. (2020). Analytic approach of response pattern of diagnostic test items in evaluating students' conceptual understanding of characteristics of particle of matter. *Journal of Baltic Science Education, 19*(5), 824-841. https://doi.org/10.33225/jbse/20.19.824

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Lewis, S. E. (2022). Considerations on validity for studies using quantitative data in chemistry education research and practice. *Chemistry Education Research and Practice, 23*, 764-767. https://doi.org/10.1039/d2rp90009b

Linacre, J. M. (2020). *A user's guide to WINSTEPS: MINISTEP Rasch-model computer programs program manual 4.5.1*. www.winsteps.com

Ling Lee, W., Chinna, K., & Sumintono, B. (2020). Psychometrics assessment of HeartQoL questionnaire: A Rasch analysis. *European Journal of Preventive Cardiology, 28*(12), e1-e5. https://doi.org/10.1177/2047487320902322

Liu, X., & Boone, W. J. (2006). *Applications of Rasch measurement in science education*. JAM Press.

Lu, S., & Bi, H. (2016). Development of a measurement instrument to assess students' electrolyte conceptual understanding. *Chemistry Education Research and Practice, 17*(4), 1030-1040.

Lu, X., & Chen, Y. (2021). Using the Rasch model to assess the psychometric properties of an online reading comprehension test for Chinese EFL learners. *Language Testing, 38*(1), 101-121. https://doi.org/10.1177/0265532220946947

Malone, K. L., Boone, W. J., Stammen, A., Schuchardt, A., Ding, L., & Sabree, Z. (2021). Construction and evaluation of an instrument to measure high school students biological content knowledge. *EURASIA Journal of Mathematics, Science and Technology Education, 17*(12), em2048. https://doi.org/10.29333/ejmste/11376

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment, 23*(2), 121-138. https://doi.org/10.1080/10627197.2018.1427570

McNeill, K. L., & Krajcik, J. (2008). Scientific explanations: Characterizing and evaluating the effects of teachers' instructional practices on student learning. *Journal of Research in Science Teaching, 45*(1), 53-78. https://doi.org/10.1002/tea.20201

Mendonça, P. C. C., & Justi, R. (2014). An instrument for analyzing arguments produced in modeling-based chemistry lessons. *Journal of Research in Science Teaching, 51*(2), 192-218. https://doi.org/10.1002/tea.21133

Ministry of Education and Culture. (2013). *Dokumen kurikulum 2013* [*2013 curriculum documents*]. Kemendikbud.

Mulford, D. R., & Robinson, W. R. (2002). An inventory for alternate conceptions among first-semester general chemistry students. *Journal of Chemical Education, 79*(6), 739. https://doi.org/10.1021/ed079p739

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

National Research Council. (2007). *Taking science to school: Learning and teaching science in grades K-8*. https://doi.org/10.17226/11625

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. https://doi.org/10.17226/13165

Neuman, W. L. (2014). *Social research methods: Qualitative and quantitative approaches*. Pearson.

Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). Towards a learning progression of energy. *Journal of Research in Science Teaching, 50*(2), 162-188. https://doi.org/10.1002/tea.21061

Nongna, C., Junpeng, P., Hong-ngam, J., Podjana, C., & Tang, K. (2023). Rasch analysis for standards-setting appraisal of competency level-based performance on the part of instructors in higher education. *Pertanika Journal of Social Science and Humanities, 31*(1), 319-338. https://doi.org/10.47836/pjssh.31.1.17

Opitz, A., Heene, M., & Fischer, F. (2017). Measuring scientific reasoning–a review of test instruments. *Educational Research and Evaluation, 23*(3-4), 78-101. https://doi.org/10.1080/13803611.2017.1338586

Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education, 95*(4), 627-638. https://doi.org/10.1002/sce.20438

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994-1020. https://doi.org/10.1002/tea.20035

Owens, D. C., Sadler, T. D., & Friedrichsen, P. (2019). Teaching practices for enactment of socio-scientific issues instruction: An instrumental case study of an experienced biology teacher. *Research in Science Education, 49*(1), 35-59. https://doi.org/10.1007/s11165-018-9799-3

Pentecost, T. C., & Barbera, J. (2013). Measuring learning gains in chemical education: A comparison of two methods. *Journal of Chemical Education, 90*(7), 839-845. https://doi.org/10.1021/ed400018v

Rahayu, S. (2019). Argumentasi ilmiah: Implementasinya dalam pembelajaran kimia untuk meningkatkan keterampilan berkomunikasi [Scientific argumentation: Its implementation in chemistry learning to improve communication skills]. In *Proceedings of the National Seminar on Chemistry 2019*.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Rickinson, M., Justin, D., Teamey, K., Morris, M., Choi, M. Y., Sanders, D., & Benefield, P. (2004). *A review of research on outdoor learning*. National Foundation for Educational Research and King's College London.

Ryder, J., & Leach, J. (1999). University science students' experiences of investigative project work and their images of science. *International Journal of Science Education, 21*(9), 945-956. https://doi.org/10.1080/095006999290246

Sadler, P. M. (1998). Psychometric models for student-conceptions in science: Reconciling qualitative studies and distractor-driver assessment instruments. *Journal of Research in Science Teaching, 35*(3), 265-296. https://doi.org/10.1002/(SICI)1098-2736(199803)35:3<265::AID-TEA3>3.0.CO;2-P

Salibašić Glamočić, D., Mešić, V., Neumann, K., Sušac, A., Boone, W. J., Aviani, I., Hasović, E., Erceg, N., Repnik, R., & Grubelnik, V. (2021). Maintaining item banks with the Rasch model: An example from wave optics. *Physical Review Physics Education Research*, 17, 010105. https://doi.org/10.1103/PhysRevPhysEducRes.17.010105

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction, 23*(1), 23-55. https://doi.org/10.1207/s1532690xci2301_2

Schwichow, M., Christoph, S., Boone, W. J., & Härtig, H. (2016). The impact of sub-skills and item content on students' skills with regard to the control-of-variables strategy. *International Journal of Science Education, 38*(2), 216-237. https://doi.org/10.1080/09500693.2015.1137651

Soeharto, S., & Csapó, B. (2021). Evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts. *Heliyon, 7*(11), E08352. https://doi.org/10.1016/j.heliyon.2021.e08352

Sovey, S., Osman, K., & Matore, M. E. E. M. (2022). Rasch analysis for disposition levels of computational thinking instrument among secondary school students. *EURASIA Journal of Mathematics Science Technology Education, 18*(3), em2088. https://doi.org/10.29333/ejmste/11794

Stevenson, C. E., Hickendorff, M., Resing, W. C. M., Heiser, W. J., & de Boeck, P. A. L. (2013). Explanatory item response modeling of children's change on a dynamic test of analogical reasoning. *Intelligence, 41*(3), 157-168. https://doi.org/10.1016/j.intell.2013.01.003

Stojanovska, M. I., Soptrajanov, B. T., & Petrusevski, V. M. (2012). Addressing misconceptions about the particulate nature of matter among secondary-school and high-school students in the Republic of Macedonia. *Creative Education, 3*(5), 619-631. https://doi.org/10.4236/ce.2012.35091

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial* [*Application of Rasch model in social sciences research*]. Trim Publishing.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi pemodelan Rasch pada assessment pendidikan* [*Application of Rasch modeling in educational assessment*]. Trim Publishing.

Szalay, L., & Tóth, Z. (2016). An inquiry-based approach of traditional "step-by-step" experiments. *Chemistry Education Research and Practice, 17*(4), 923-961. https://doi.org/10.1039/c6rp00044d

Taber, K. S. (2002). *Chemical misconceptions – Prevention, diagnosis, and cure*. Royal Society of Chemistry.

Taber, K. S. (2009). Challenging misconceptions in the chemistry classroom: Resources to support teachers. *Educació Química* [*Chemical Education*], *4*, 13-20.

Taber, K. S. (2014). Ethical considerations of chemistry education research involving "human subjects". *Chemistry Education Research and Practice, 15*(2), 109-113. https://doi.org/10.1039/c4rp90003k

Talanquer, V. (2018). Progressions in reasoning about structure-property relationships. *Chemistry Education Research and Practice, 19*(4), 998-1009. https://doi.org/10.1039/c7rp00187h

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159-169. https://doi.org/10.1080/0950069880100204

Tseng, W. T., Su, T. Y., & Nix, J. M. L. (2019). Validating translation test items via the many-facet Rasch model. *Psychological Reports, 122*(2), 748-772. https://doi.org/10.1177/0033294118768664

Uzuntiryaki, E., & Aydin, Y.C. (2009). Development and validation of chemistry self-efficacy scale for college students. *Research in Science Education, 39*(4), 539-551. https://doi.org/10.1007/s11165-008-9093-x

Van Vo, D., & Csapó, B. (2021). Development of scientific reasoning test measuring control of variables strategy in physics for high school students: evidence of validity and latent predictors of item difficulty. *International Journal of Science Education, 43*(13), 2185-2205. https://doi.org/10.1080/09500693.2021.1957515

Wang, C. Y. (2015). Scaffolding middle school students' construction of scientific explanations: Comparing a cognitive versus a metacognitive evaluation approach. *International Journal of Science Education, 37*(2), 237-271. https://doi.org/10.1080/09500693.2014.979378

Wang, W. C., & Willson, M. (2005). Exploring local item dependence using a random-effects facet model. *Applied Psychological Measurement, 29*(4), 296-318. https://doi.org/10.1177/0146621605276281

Wei, S., Liu, X., Wang, Z., & Wang, X. (2012). Using rasch measurement to develop a computer modeling-based instrument to assess students' conceptual understanding of matter. *Journal of Chemical Education, 89*(3), 335-345. https://doi.org/10.1021/ed100852t

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9781410611697

Wilson, M. (2008). Cognitive diagnosis using item response models. *Zeitschrift Für Psychologie* [*Journal of Psychology*], *216*(2), 74-88. https://doi.org/10.1027/0044-3409.216.2.74

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716-730. https://doi.org/10.1002/tea.20318

Wind, S. A., & Gale, J. D. (2015). Diagnostic opportunities using Rasch measurement in the context of a misconceptions-based physical science assessment. *Science Education, 99*(4), 721-741. https://doi.org/10.1002/sce.21172

Wind, S. A., Tsai, C. L., Grajeda, S. B., & Bergin, C. (2018). Principals' use of rating scale categories in classroom observations for teacher evaluation. *School Effectiveness and School Improvement, 29*(3), 485-510. https://doi.org/10.1080/09243453.2018.1470989

Wu, H. K., & Hsieh, C. E. (2006). Developing sixth graders' inquiry skills to construct explanations in inquiry-based learning environments. *International Journal of Science Education, 28*(11), 1289-1313. https://doi.org/10.1080/09500690600621035

Yang, J., Chang, H. H., & Cheng, Y. (2021). Current trends in Rasch modeling in educational research: A systematic review of the literature. *Educational Research Review, 33*, 100406. https://doi.org/10.1016/j.edurev.2021.100406

Yao, J. X., & Guo, Y. Y. (2018). Validity evidence for a learning progression of scientific explanation. *Journal of Research in Science Teaching, 55*(2), 299-317. https://doi.org/10.1002/tea.21420

Yasar, I. Z., Ince, E., & Kirbaslar, F. G. (2014). 7. class science and technology course "structure of atom" subject readiness improvement test. *Procedia-Social and Behavioral Sciences, 152*, 662-667. https://doi.org/10.1016/j.sbspro.2014.09.259

Yildirir, H. E., & Demirkol, H. (2018). Identifying mental models of students for physical and chemical change. *Journal of Baltic Science Education, 17*(6), 986-1004. https://doi.org/10.33225/jbse/18.17.986

Zhan, P., Jiao, H., & Liao, D. (2017). Cognitive diagnosis modelling incorporating item response times. *British Journal of Mathematical and Statistical Psychology, 70*(2), 332-355. https://doi.org/10.1111/bmsp.12114

APPENDIX A

Table A1. Measure of test items & distribution of total students on level of ability in constructing scientific explanations on chemical phenomena

Item	M	Achieved MPI abilities (%)				Chemical phenomena	Basic concept of understanding
		HACSE	MACSE	LACSE	NACSE		
B18	.71	4	15	69	13	Process of hydrolysis	Salt hydrolysis
C25	.55	11	17	54	18	Electroplating on metal	Elemental chemistry
B7	.46	17	9	61	14	Petroleum formation	Petroleum
B5	.45	12	14	65	9	Coal formation	Hydrocarbon
B20	.45	17	9	60	14	Blood pH regulation	Buffer solution
B11	.42	16	8	69	8	Melting ice cubes	Thermochemistry
B9	.36	18	8	66	8	Process of photosynthesis	Thermochemistry
B16	.33	20	13	52	15	Use of fertilizer	Salt hydrolysis
B19	.29	23	10	51	15	Process of weathering	Salt hydrolysis
B10	.28	24	13	44	19	Water evaporation	Thermochemistry
C26	.26	18	18	54	10	Fireworks	Elemental chemistry
C28	.23	21	18	47	14	Food preservatives	Benzene & its derivatives
B22	.21	23	13	52	12	Process of salt solubility in water	Solubility & solubility product
B8	.18	18	23	49	10	Wood burning	Thermochemistry
B14	.16	25	9	57	9	Antacids for stomach ulcers	Acid & base
B13	.14	24	22	36	17	CO_2 formation from baking soda & vinegar	Acid & base
B6	.08	28	17	39	16	Garbage decomposition	Redox
C29	.08	27	16	44	12	Process of fermentation	Macromolecule
A3	-.11	35	16	36	13	Color changes in apple	Redox
B21	-.22	34	23	34	9	Deficiency in red blood cells in body	Buffer solution
A1	-.27	43	9	40	8	Rusting iron	Redox
A4	-.33	37	26	30	8	Rotting banana	Redox
A2	-.40	48	19	15	18	Fruit rot	Redox
B24	-.40	49	11	29	11	Use of sunscreen	Colloid
C27	-.45	44	23	24	10	Firework flaming colors	Elemental chemistry
C26	-.48	53	8	29	10	Fireworks	Elemental chemistry
B17	-.57	50	22	18	10	Bleach on cloth	Salt hydrolysis
B23	-.72	60	10	22	8	Drinking water purification process	Colloid
B12	-.72	50	25	21	4	Process of acid rain	Acid & base
B15	-.96	64	16	12	7	Use of detergent	Salt hydrolysis

Note. M: Measure; HACSE: High ability in constructing scientific explanations; MACSE: Moderate ability in constructing scientific explanations; LACSE: Low ability in constructing scientific explanations; & NACSE: No ability in constructing scientific explanations

APPENDIX B

Table B1. Measure of test items of ability in constructing scientific explanations on chemical phenomena on students' different classes

Item	M	Students' classes/item measure (logit)									
		A	B	C	D	E	F	G	H	I	J
A1	-.27	-.03	-.48	.45	-.39	-.40	-.58	-.65	-.11	.15	-.04
A2	-.40	-.62	-.12	-.30	-.51	-.38	-.31	-.83	-.32	-.28	-.27
A3	-.11	-.07	-.42	-1.37	-.45	-.35	.10	-.10	-.14	-.17	.22
A4	-.33	-.53	-.08	.78	-.51	-.09	-.39	-.36	-.51	-.33	-.29
B5	.45	.55	-.06	.91	.30	.59	.20	.51	.35	.99	.90
B6	.08	-.04	.31	-.71	.07	.16	.35	.07	.23	.11	-.58
B7	.47	.58	.38	-.30	.47	.24	.42	.74	0	.67	.56
B8	.18	.43	.05	.91	.07	-.04	.01	.16	.15	.29	.11
B9	.36	.86	.53	-.21	.74	.82	-.04	.25	-.02	.37	-.02
B10	.28	.27	.11	-.60	.30	.62	.57	.23	.44	.27	-.11
B11	.42	.40	.23	.35	.07	.56	.14	.60	.47	.61	1.15
B12	-.72	-.78	-1.24	-.12	-.33	-.14	-.86	-.58	-.62	-.82	-.67
B13	.14	.05	-.02	.25	-.20	-.09	.52	.43	.17	-.42	.26
B14	.16	-.09	.49	-.6	.47	.01	.16	.54	-.11	.17	.50
B15	-.96	-1.07	-.84	-1.07	-1.29	-1.08	-1.01	-1.23	-.58	-.75	-.83
B16	.33	.37	-.17	.25	.84	.24	.44	.41	.03	.61	.47
B17	-.57	-.59	-.54	.35	-.51	-.82	-.85	-.60	-.44	-.45	-.20
B18	.71	.79	.65	1.35	.74	.62	.6	.92	.72	.99	.52
B19	.29	.32	.09	-.03	.74	.32	.51	.51	.29	.07	0
B20	.45	.82	.36	1.35	.55	.24	.49	.67	.23	.15	-.11
B21	-.22	-.39	.05	.67	-.07	-.09	-.35	-.31	.20	-.63	-.13
B22	.21	.35	.44	.16	.07	.27	.23	-.01	.12	.09	-.08
B23	-.72	-.84	-.57	-.71	-.95	-.96	-.88	-.60	-.20	-.75	-.61
B24	-.40	-.71	-.02	-.94	-.13	-.59	-.11	-.49	-.95	-.28	-.64
C25	.55	.48	.35	.91	.38	.59	.57	.54	.94	.65	.70
C26	.26	-.56	-.45	-.94	-.13	-.57	-.46	-.75	-.55	-.37	-.20
C27	.26	.50	.51	-.03	0	-.09	.06	.18	.09	.13	.59
C28	-.45	-.76	-.42	-.82	-.33	-.50	-.17	-.08	-.38	-.92	-.73
C29	.23	.32	.46	.16	-.33	.38	.52	-.15	.29	-.09	-.13
C30	.08	-.03	.40	-.12	.30	.50	.13	-.02	.20	-.09	-.36