Jurnal Pendidikan Sains Indonesia 11(2), p.360-378 (2023)
(Indonesian Journal of Science Education) e-ISSN: 2615-840X p-ISSN: 2338-4379
URL: http://jurnal.unsyiah.ac.id/JPSI/index DOI: doi.org/**10.24815/jpsi.v10i4.27822**

Measuring Students' Scientific Argumentation Skills in Explaining Phenomena Related to Acid-Base Concepts

Rahmawati Hilala, Lukman Abdul Rauf Laliyo*, Jafar La Kilo, Julhim Tangio, Erni Mohammad, Mangara Sihaloho

Program Studi Pendidikan Kimia FMIPA Universitas Negeri Gorontalo, Jl. Prof. Dr. B.J. Habibie, Tilongkabila, Gorontalo, Indonesia

*Email: lukmanlaliyo2020@gmail.com

Article History:
Received date: August 29, 2023
Received in revised from: March 2, 2023
Accepted date: April 11, 2023
Available online: April 27, 2023

Citation:
Hilala, R., Laliyo, L.A.R., Kilo, J.L., Tangio, J., Mohammad, E., & Sihaloho, M. 2023. Scientific argumentation skills in explaining phenomena related to acid-base concepts. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, 11(2):360-378.

Abstract. Scientific argumentation is one of the main competencies of students in communicating chemical phenomena through the application of conceptual mastery that they have understood. The development of this ability has been postulated in the 2013 curriculum through constructivist-based learning approaches and scientific inquiry, either in the classroom or the laboratory. However, there is relatively little information on how students' scientific argumentation skills develop. This study aims to measure students' scientific argumentation skills in explaining five science phenomena related to the acid-base concept, namely: acid rain, salt crystal making, the use of antacids as ulcer medicine, the use of $CaMgCO_3$ dolomite fertilizer, and the difference in the acidity level of HCl and H_2SO_4, using Rasch modeling. Each phenomenon was measured by three multiple-choice test items, which were developed to test students' ability to make claims (Q1), evidence (Q2), and justification (Q3). The data were analyzed using Rasch modeling, allowing researchers to measure the item and individual respondent levels. Respondents were 100 chemistry students in Gorontalo, who were differentiated in gender and adversity quotient. The results showed that the measurement instrument has good validity and reliability. In addition, it was found that students' abilities differed; some items responded differently regarding gender and adversity quotient. Most students tended to be weak in explaining the phenomenon of acid rain, the use of antacids for ulcer disease, the use of dolomite fertilizer, and the difference in the acidity level of HCl and H_2SO_4.
Keywords: Rasch model; argumentation; adversity quotient; gender

Introduction

Scientific argumentation is one of the important competencies in scientific inquiry. It is one of the cognitive skills that tend to be complex, requiring reasoning between theory and evidence and critical thinking (Mao et al., 2018). For students, this competency can be reflected in their ability to explain scientific phenomena based on evidence/data, theory, and valid reasoning to support their explanation (Rahayu, 2017; Samosa, 2021). A claim is a descriptive statement to answer a problem, while evidence/data is information collected, analyzed, and interpreted to support a claim. Further, justification is determined through valid reasoning supported by principles, concepts, laws, or assumptions. This justification is used to explain the relevance of evidence/data and claims (Muntholib et al., 2021; Rahayu, 2019; Talanquer, 2018).

The ability of students to build scientific arguments regarding chemical phenomena, including the occurrence of acid rain, the use of antacids as ulcer medicine, etc., is an example where this phenomenon can be understood if students have a good understanding of the acid-base concept. In this context, the level of mastery of students' abilities can be referred to as their explanation of the question of how and why the phenomenon can occur based on three main components: claims, evidence, and reasoning (Chin & Brown, 2000). A claim is an initial, brief answer containing agreement, refutation, substance, or classification. Submission of evidence is an argument related to examples or data that can support a claim, which students understand through their learning experience. The reasoning is an argument that connects claims and evidence, supported by the use of images, graphs, and mathematical formulas (McNeill & Krajcik, 2008).

In Indonesia, the ability to build scientific explanations has been explicitly postulated in Curriculum 2013, through the implementation of constructivist-based learning and scientific inquiry, both in the laboratory and in the classroom, evenly throughout secondary schools in Indonesia (Rahayu, 2019). However, as far as researchers have searched, it is relatively difficult to find scientific information which shows the effectiveness of learning in developing students' scientific argumentation skills. Chemistry education researchers, especially in Indonesia, tend not to concentrate much on measuring students' scientific argumentation. This scenario may have led to limited scientific information regarding the effectiveness of chemistry learning developed based on the 2013 Curriculum. In addition, this limitation is also thought to be constrained by the limited measurement and analysis techniques used.

Developing the ability to build scientific argumentation tends to be related to student learning experiences (Shin et al., 2019; Testa et al., 2019). Previous research has shown that learner engagement in the learning process and scientific practice has distinctive and epistemic characteristics (Deng & Wang, 2017; Driver et al., 2000; Erduran et al., 2004; Osborne et al., 2004). Therefore, in recent decades, experts have developed a variety of new methods and curricula to help develop learners' ability to construct scientific arguments through meaningful learning engagement (Hong et al., 2013). This has led to increased research on evaluating learners' ability to construct scientific argumentation (Mendonça & Justi, 2014; Sandoval & Millwood, 2005). Unfortunately, research in the field of chemistry education that evaluates students' ability to build scientific argumentation tends to be rare (Cetin, 2014; Rahayu, 2019).

Some previous studies reported measuring scientific argumentation skills, where students were asked to submit various knowledge claims on certain content measures and develop their reasoning to evaluate their scientific argumentation framework (Ural & Gencoglan, 2019). According to Erduran et al. (2004), this is a way of measuring students' ability to construct, support, evaluate, or validate claims with evidence-based reasoning. In addition, students' scientific argumentation skills are tested based on their ability to express scientific theories, data and evidence to confirm or refute claims (Aydeniz et al., 2012; Sandoval, 2003). Recent research developments show that students' scientific argumentation skills are seen from their ability to develop scientific argumentation processes to explain scientific phenomena using concept mastery. It is also supported by data in the form of examples or observations, which can be justified in principle through the relationship between claims and evidence (Tümay, 2016).

The focus of this research is intended to measure the ability to build scientific argumentation, which is assumed to have been developed by students based on their learning experiences since high school, based on the 2013 Curriculum, to a formal study of chemistry in college. In the context of this study, students' scientific argumentation skills are evaluated in terms of their ability to explain five scientific phenomena, including (the phenomenon of acid rain, the manufacture of salt crystals, the use of antacids as ulcer medicine, the use of $CaMgCO_3$ dolomite fertilizer, and the difference in the acidity level of

HCl and H_2SO_4). The scientific argumentation of the five phenomena requires the ability to master the acid-base concept. The assumption is that the better the mastery of acid-base concepts, the better the ability to build scientific argumentation of students. Acid and base materials are the conceptual footings of students in developing their scientific argumentation (Damanhuri et al., 2016).

Several studies have reported that, basically, students tend to have weaknesses in understanding acid-base concepts due to the existence of (Bradley & Mosimege, 1998; McClary & Bretz, 2012). Does the weakness in mastering the acid-base concept tend to affect students' ability to develop scientific argumentation of a scientific phenomenon? Answering this question requires valid and reliable measurements. For this purpose, an instrument was developed that aims to measure students' ability to build scientific arguments for five science phenomena, namely: (1) the occurrence of acid rain, (2) the manufacture of salt crystals, (3) the use of antacids as ulcer medicine, (4) the use of $CaMgCO_3$ dolomite fertilizer, and (5) the difference in the acidity level of HCl and H_2SO_4. The instrument was designed in the form of a multiple-choice diagnostic test. Each item contains three graded questions (Q1, Q2, and Q3). Question Q1 measures students' ability to make claims. Question Q2 measures students' ability to provide evidence, and question Q3 measures students' ability to justify/reason. The design of this test instrument was adapted from the development of a two-tier multiple-choice test recommended by Treagust (1988) and Chandrasegaran et al. (2007) and added one question (Q3) that tests students' ability to connect understanding claims (Q1) with evidence/data (Q2). Five answer options were provided for each layer of questions, one correct answer option, three distractor answer options, and one other answer option. These distractor answer choices are answer choices that contain misconceptions. The function of distractor answer choices is to increase the diagnostic power of the item (Herrmann-Abell & Deboer, 2016; Herrmann-Abell & DeBoer, 2011; Sadler, 1999).

The use of multiple-choice diagnostic item-based instruments in measuring students' scientific argumentation skills tends not to be widely used. Most diagnostic test instruments tend to be used to diagnose misconceptions. For example, research reported by Lu & Bi (2016) on identifying students' preconceptions in understanding the concept of electrolytes. Similarly, research reported by Luxford & Bretz (2014) related to misconceptions in understanding covalent bonding and chemical equilibrium (Tyson et al., 1999). In this study, a diagnostic item test was used to measure students' abilities on each question (Q1, Q2, and Q3) separately for each phenomenon. This was intended to diagnose barriers in students' ability to develop accurate and scientific argumentation (Gabel, 1999; Gette et al., 2018), which can be found in each question (Q1, Q2, Q3).

Analysis of student responses to each question used the Rasch model. One of the advantages of this analytical approach is that it does not use raw scores, but converts them into interval data with the same scale, making it more accurate in diagnosing student misconceptions up to the level of individual items and individual students (Sumintono & Widhiarso, 2015). In contrast to other misconception research, it tends to be based on raw scores, even though raw scores are not final data, so they do not have much information to make conclusions (He et al., 2016; Sumintono, 2016). Research conclusions built on raw scores tend to provide limited information in explaining item and individual learner misconceptions (Pentecost & Barbera, 2013). Using Rasch modeling tends to serve to overcome the weaknesses of conventional raw score-based psychometric measurements (Perera et al., 2018); a probabilistic measurement model goes beyond the use of raw scores (Boone & Staver, 2020; Wei et al., 2012).

In addition to measuring students' ability to build scientific argumentation, this study also measured whether or not items were responded to differently regarding differences in adversity quotient (AQ) and student gender. The adversity quotient is defined as the ability of students to face a problem and find a solution to a problem. Characterizing students' adversity quotient (AQ) tendencies are intended to determine whether students easily give up in overcoming difficulties or, on the contrary, never give up and always try to overcome their difficulties (Muhammad, 2020); likewise, differences in student gender. There are two questions explained in this study, namely: (1) To what extent is the effectiveness of this research instrument, in terms of the validity and reliability of the instrument, when used to measure students' scientific argumentation skills in explaining scientific phenomena that are conceptually bound to the concept of acid-base? (2) How are the differences in students' scientific argumentation skills in explaining scientific phenomena that are conceptually bound to the concept of acid-base in terms of gender differences and adversity quotient?

Methods

This quantitative study used a non-experimental design, where students' scientific argumentation ability in explaining five science phenomena related to the acid-base concept was considered as a variable that could be measured. There are fifteen items of this test developed in the form of a multiple-choice diagnostic test. The measurement results are in the form of numbers, while scores are obtained for correct answers to the items. Numerical numbers represent abstract concepts that are measured empirically (Chan et al., 2021; Neuman, 2014). The present study did not make any interventions to the learning process or materials. In other words, no treatment was given to students so that they were able to answer all the items on the measurement instrument. Permission to conduct the research was granted by the head of the university.

The respondents of this study were 100 chemistry students from one of the universities in Gorontalo. The age of the respondents was 19-21 years old. With regard to research principles and ethics, students who participated voluntarily in this study have been asked for their consent. Their identities were kept confidential, and the information obtained was only intended for scientific development (Taber, 2014). The study later grouped the students based on adversity quotient (AQ) ability, using a questionnaire containing 40 statements to be filled out by students. The questionnaire contains a scale, which is divided into two statements, namely positive and negative. The adversity quotient scale refers to aspects of the adversity quotient: control, ownership, reach, and endurance (Stoltz, 2018). Each statement consists of four answer categories: very appropriate, appropriate, inappropriate, and very inappropriate (Hidayat, 2017). The demographic profile of the respondents is shown in Table 1.

Table 1. Demographic profile of respondents

Demographic		Respondent	Respondent (%)
Adversity quotient			
	High	17	0,17
	Medium	68	0,68
	Low	15	0,15
Gender			
	Male	14	0,14
	Female	86	0,86

The stages of instrument development in this study followed the recommendations of Wilson (2005) and Wilson (2008). The instrument consisted of fifteen multiple-choice test items. The first stage was to determine the constructs from the construct map. Table 2 presents the construct map that presents the definition of the measured construct, namely the scientific argumentation ability of students in explaining five science phenomena related to the acid-base concept.

Table 2. Constructing a map of the scientific argumentation ability of five science phenomena related to the concept of acid-base

Science phenomena	Indicators of scientific argumentation skills	Item
Acid rain phenomenon	• Students are able to provide claims related to statements about the effect of acids and bases on the environment.	A-Q1
	• Students are able to provide evidence related to the phenomenon of acid rain	A-Q2
	• Students are able to provide justification related to the claims and evidence of the acid rain phenomenon	A-Q3
Manufacture of salt crystals	• Students are able to make claims related to statements about acid anions and base anions that produce (neutral) salts	B-Q1
	• Students are able to provide evidence related to the manufacture of salt crystals	B-Q2
	• Students provide justification related to the claims and evidence of the manufacture of salt crystals	B-Q3
The use of antacids as ulcer drugs	• Students are able to provide claims related to statements about acid-base neutralization in daily life, namely in ulcer medicine.	C-Q1
	• Students are able to provide evidence related to the use of antacids as ulcer drugs	C-Q2
	• Students are able to provide justification related to claims and evidence of the use of antacids as ulcer drugs	C-Q3
Use of $CaMgCO_3$ dolomite fertilizer	• Students are able to provide claims related to statements about the nature of acid-base in everyday life	D-Q1
	• Students are able to provide evidence related to the answer of Q1	D-Q2
	• Students are able to provide justification related to Q1 and Q2	D-Q3
The difference in acidity level of HCl and H_2SO_4 solutions	• Students are able to provide claims related to the statement about the acidity level of a solution	E-Q1
	• Students are able to provide evidence related to the difference in the acidity level of HCl and H_2SO_4 solutions	E-Q2
	• Students are able to provide justification related to claims and evidence. The difference in the acidity level of HCl and H_2SO_4 solutions	E-Q3

The second stage is item and assessment design. This stage is concerned with determining the types of items used to obtain evidence of students' level of construct understanding in relation to the construct map (Wilson, 2009). Each science phenomenon was measured with three items (Q1, Q2, Q3). Each item provided five answer options, consisting of one correct answer option, three answer options that are distractors, and one other answer option that students can fill in when the available answers are considered incorrect. The correct student response to each item is given a score of 1; the wrong response is given a score of 0. An example of item design is presented in Table 3.

Table 3. Example of item design B-Q1, B-Q2, and B-Q3 to measure students' scientific argumentation ability in explaining the science phenomenon of salt crystal making.

Item Description:
In everyday life, we are certainly familiar with salt. An example of salt is table salt which is commonly used for cooking purposes. Table salt can be obtained from seawater. Salt farmers make it by evaporation and crystallization. Salt is also obtained by mixing acids and bases.
Item B-Q1 (Claim). According to you, when acids and bases are mixed in the process of making salt, the properties of the salt produced are:
a. salt that is neutral (pH = 7)
b. salt that is acidic or basic
c. Salts that are amphoteric or neither acidic nor basic
d. Salts that are not neutral
e. Other answers.
Item B-Q2 (Evidence). This is the correct acid-base chemical reaction process listed below as evidence of your answer choice in the above item (B-Q1).
a. $NaOH + HCl \rightarrow NaCl + H_2O$
b. $NH_4^+ + Cl^- \rightarrow NH_4Cl$
c. $Na^+ + CH_3COONa \rightarrow CH_3COONa$
d. $NH_4^+ + H_2O \rightarrow NH_3 + H_3O^+$
e. Other answers.
Item B-Q3 (Justification). Based on your answer choices in items B-Q1 and B-Q2, the salt that results from a completely hydrolyzed acid anion and the base cation is:
a. Salts consisting of strong acid anions and weak base cations produce neutral solutions because the ions do not react with water.
b. Salts consisting of strong acid anions and strong base cations produce neutral solutions because the ions do not react with water
c. Salts consisting of weak acid anions and strong base cations produce neutral solutions because the ions do not react with water
d. Salts consisting of weak acid anions and weak base cations produce neutral solutions because the ions do not react with water
e. Other answers.

The third stage is the Rasch model measurement. This analysis approach combines the probabilistic expected outcome algorithms of item 'i' and student 'n', as: Pni (X_{ni} = $1/(\beta n, \delta i$) = (e^(($\beta n - \delta$_i)))/(1 + ($\beta n - \delta$_i)). The statement P_ni (X_{ni} = $1/(\beta n, \delta i$) is the probability of learner n in item i, to produce a correct answer (x = 1); with learner ability, ßn, and item difficulty δi (Bond & Fox, 2015). The above equation can be simplified by inserting a logarithmic function, becoming log (Pni (X_{ni} = $1/(\beta n, \delta i$)) = $\beta n - \delta i$, so the probability of success is equal to the student's ability minus the item difficulty. Learner (person) size βn and item δi are expressed on the same interval scale and are mutually independent. Learner ability and item difficulty are measured in logarithmic units called odds or logs, which can vary from -00 to +00 (Boone et al., 2014; Boone & Staver, 2020; Sumintono, 2014).

Further, Winsteps modeling software version 4.5.5 was used to convert the raw data into interval data. The conversion results in the calibration of student ability level data and item difficulty levels in the same interval size. The analysis of the diagnostic test item response pattern is carried out in two stages: First, converting the raw score into the same interval size while analyzing the effectiveness of the measurement instrument. Second, measuring differences in argumentation skills using the differential item functioning (DIF) item test.

Results and Discussion

One of the important requirements of Rasch modeling testing is that the instruments used meet the requirements of instrument validity and reliability (Boone & Staver, 2020). The test is intended to determine how far the instruments used in this study produce consistent information in revealing latent traits or the unidimensional nature of the measured variables (Sumintono & Widhiarso, 2015). In this study, the instrument effectiveness requirements were tested based on the parameters and summary of fit statistics. The results are presented in Table 4.

Table 4. Summary of fit statistics

Parameter	INFIT		OUTFIT		Separation	Reliability	Measure	KR-20	SD
(N)	MNSQ	ZSTD	OUTFIT	ZSTD					
Person (100)	.99	.08	.98	.1	1.35	.65	.37	.65	.87
Item (15)	1.02	-.08	.98	-.19	2.53	.86	0.0		.61

From Table 4, it is known that the person reliability value is .65 logit which is equivalent to a person separation index value of 1.35 logit. This means that the consistency of student responses to the test is fairly weak. Similarly, the Cronbach alpha coefficient (KR-20) value of .65 indicates the interaction of students with the test, which is considered rather weak. In addition, an item separation index value of 2.53 was obtained, which is equivalent to an item reliability value of .86. It means that item consistency is good or items can be said to meet the requirements of unidimensionality (Bond & Fox, 2015).

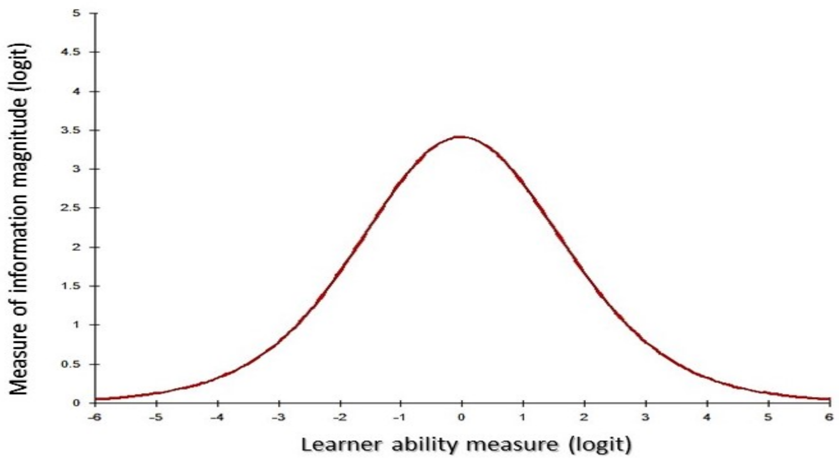

Figure 1. Measurement information function

Figure 1 presents the measurement information graph. This graph shows the trend of measurement reliability. The higher the peak of the information function graph, the higher the measurement reliability value. At the student ability level between the range of

-4.0 logit to +4.0 logit), the measurement information is quite high. This shows that the instrument used is able to produce optimal information on medium-ability students. This means the instrument has good reliability in measurement (Laliyo et al., 2022; Sumintono & Widhiarso, 2015).

Item validity testing is used to ensure that all items fit the Rasch model, in this case, tested with the Fit item parameter. The goal is to find out whether the test items measure what should be measured, or called test validity, based on three criteria, namely: outfit means square (MNSQ): $0.5 < y < 1.5$; outfit z-standard: $-2.00 < Z < +2.0$, and point measure (PTMEA Corr): $0.4 < x < 0.8$, where the value must be positive and not close to zero (Bond & Fox, 2015). If all three criteria are not met, then the item is not good enough and needs further research (Boone et al., 2014). Both Outfit MNSQ and Infit MNSQ are sensitive chi-squares to detect outlier response patterns. There are two types of outlier responses: correct responses by guessing by low-ability students on items with high difficulty levels; or incorrect responses due to carelessness by high-ability students on items with low difficulty levels (easy). The expected ideal value of MNSQ is 1.0. The results of the item suitability analysis are presented in Table 5. From this table, it is known that all items meet the MNSQ outfit criteria, and there is no PT MEA Corr with a negative value. This means that there are no deviant items; the items are appropriate and valid. Although some items do not meet one of the criteria, it does not reduce the quality of the items.

Table 5. Item statistics: misfit order

No	Item	Measure	Outfit MNSQ	Outfit ZSTD	PTMEA Corr
11	D-Q2	1.53	1.45	1.43	.06
3	A-Q3	-.14	1.10	.75	.30
2	A-Q2	.21	1.14	.97	.39
13	E-Q1	.21	1.09	.65	.44
15	E-Q3	.38	1.08	.57	.36
12	D-Q3	.21	1.04	.20	.37
5	B-Q2	-1.16	.99	.02	.36
10	D-Q1	-.48	.96	-.26	.40
14	E-Q2	.38	.94	-.32	.44
1	A-Q1	.72	.90	-.50	.46
4	B-Q1	-.86	.93	-.39	.40
6	B-Q3	-.72	.84	-1.01	.45
9	C-Q3	-.48	.83	-1.23	.49
7	C-Q1	-.19	.76	-1.90	.56
8	C-Q2	.38	.71	-1.98	.61

Figure 2 is a Wright Map illustrating the distribution of items along the line of student ability levels. This map is used to measure the consistency of item difficulty and student ability level. The higher the item difficulty, the higher the students' abilities (Boone, 2016). Wright map information: person-map-item is presented in Figure 2. From Wright's map, it is known that all instrument items can cover most of the student's abilities. There are students with the high ability (>2.0 logits) and very low ability (<-2.0 logits). The location of item difficulty levels is mostly in the interval (-1.0 logit to +1.0 logit), even though some items tend to be in the same item difficulty level. Item (D-Q2) is the item with the most serious difficulty level (+1.53 logits), and item B-Q2 is the item with the lowest difficulty level (-1.16 logits).

In terms of the difference in item size, several cases can be explained as follows: First, item size D-Q2 (+1.53) > D-Q3 (+0.21) > D-Q1 (-0.48), where D-Q2 has the most serious difficulty. This shows that for the context of the problem of using $CaMgCO_3$ dolomite fertilizer, students are more likely to find it difficult to provide evidence than claims and justifications. Second, item size A-Q1 > A-Q2 > A-Q3. This shows that in the context of the acid rain problem, students tend to find it difficult to provide claims, evidence, and justifications that are in line. Then seen from the size of item C-Q2 > C-Q1 > C-Q3, showing students tend to find it difficult to provide evidence, claims, and justifications for the context of the problem of solving the problem of using antacids for ulcer medicine. Furthermore, item size E-Q2 > E-Q3 > E-Q1, for the context of the problem of differences in the acidity of HCl and H_2SO_4, shows that students tend to find it difficult to provide evidence, then justification and claims. For the context of the problem of making salt crystals, item size B-Q3 > B-Q1 > B-Q2 shows that students tend to find it difficult to provide justification than claims and evidence. Various differences in the size of items Q1, Q2, and Q3 in each problem context, which is not sequential according to the construct level, where the item size should be Q1 < Q2 < Q3, indicate that students tend not to be able to provide good and correct scientific argumentation. Students' understanding of the problem context is not in line with the claims, evidence, and justifications they provide (Probosari et al., 2016).

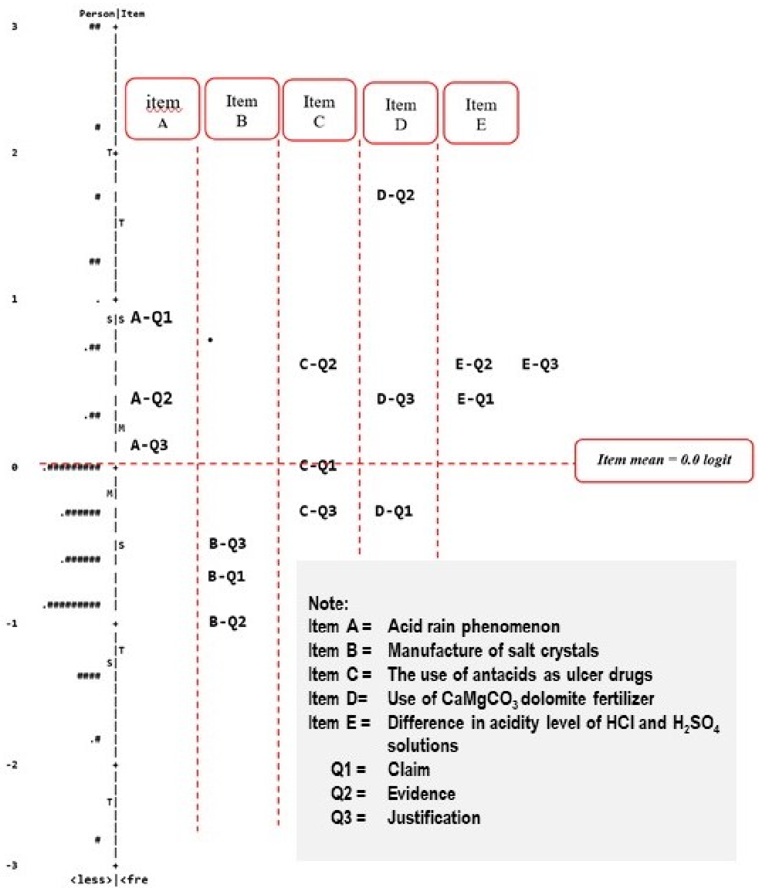

Figure 2. Person-item wright map

Differences in scientific argumentation skills based on gender differences

Differences in students' argumentation skills related to the concept of acid-base based on gender were analyzed by differential item functioning (DIF). Figure 3 presents the DIF plot of the scientific argumentation ability of male and female students. From this figure, the following facts can be explained: First, in terms of the ability to provide claims (Q1), it was found that male students have the ability to provide claims (Q1), easier on items D-Q1>E-Q1>C-Q1>B-Q1>A-Q1; while female students are easier on items A-Q1>E-Q1>C-Q1>D-Q1>B-Q1. This finding/fact shows that male students most easily give claims in the context of problems related to the phenomenon of acid rain (item A-Q1), and female students most easily give claims in the context of problems related to the manufacture of salt crystals (item B-Q1). However, male students had the most difficulty making claims in the problem context of using $CaMgCO_3$ dolomite fertilizer (item D-Q1), and female students had the most difficulty making claims in the problem context of the acid rain phenomenon (item A-Q3).

Second, in terms of the ability to provide evidence (Q2), it was found that male students had an easier ability on item D-Q2>C-Q2>E-Q2>A-Q2>B-Q2; while female students had an easier time on item D-Q2>E-Q2>A-Q2>C-Q2>B-Q2. This means that both male and female students most easily provide evidence in the context of problems related to making salt crystals (item B-Q2). However, both had the most difficulty providing evidence in the context of the item on the use of $CaMgCO_3$ dolomite fertilizer (item D-Q2).

Third, in terms of the ability to provide justification (Q3), it is found that male students have the ability to provide claims (Q3) more easily on items E-Q3>D-Q3>C-Q3>C-Q3>B-Q3, while female students on items E-Q3>D-Q3>A-Q3>C-Q3>B-Q3. This means that male and female students most easily provide justification in the context of problems related to the manufacture of salt crystals (item B-Q3). However, both of them had difficulty in providing justification in the context of the item on the difference in the acidity level of HCl and H_2SO_4 (item E-Q3).

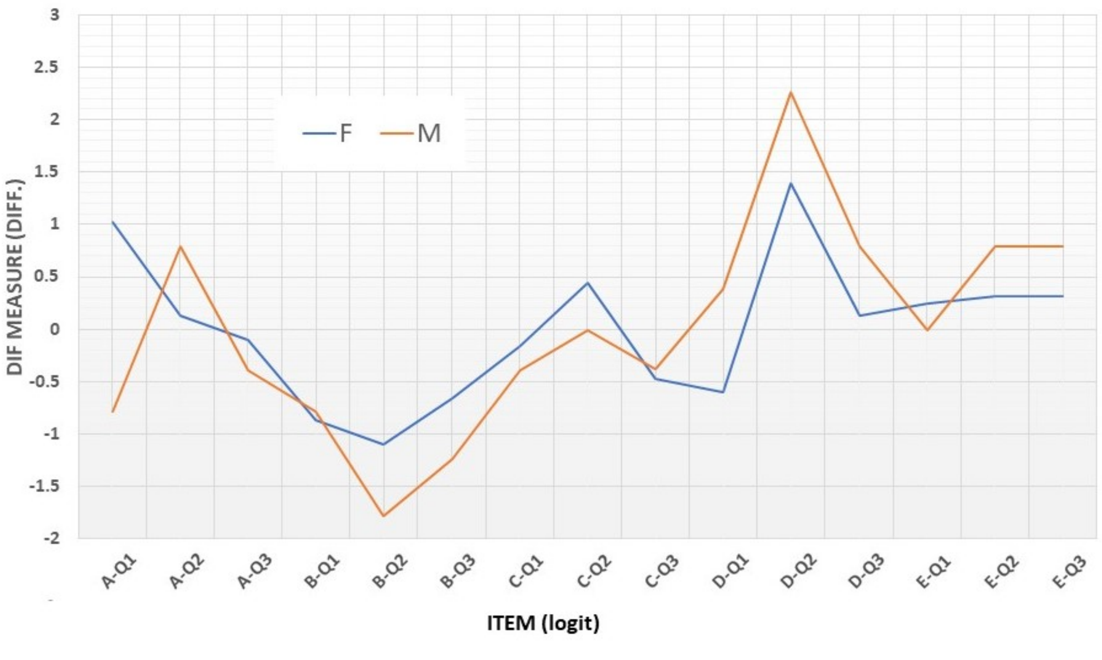

Figure 3. DIF plot of students' scientific argumentation ability based on gender differences. M= male students, F= female students

Differences in scientific argumentation skills are based on differences in adversity quotient (AQ)

Figure 4 shows the DIF plot based on the scientific argumentation ability of students who have an AQ index (high, medium, and low). In terms of this difference in AQ ability, several interesting facts can be explained. First, the difference in scientific argumentation ability in providing claims (Q1). High AQ students, giving claims item E-Q1>A-Q1>C-Q1>D-Q1>B-Q1, medium AQ students item A-Q1>E-Q1>C-Q1>E-Q1>D-Q1. Low AQ students, item A-Q1>E-Q1>C-Q1>E-Q1>D-Q1. This finding shows that high AQ students tend to give the easiest claims in the context of the problem of making salt crystals (item B-Q1), while medium and low AQ students in the context of problems related to the use of $CaMgCO_3$ dolomite fertilizer (D-Q1). However, the most difficult thing to do is to give claims in the context of items on the difference in the acidity level of HCl and H_2SO_4 (item E-Q1) and the phenomenon of acid rain (item A-Q1).

Second, high AQ students have the ability to provide evidence (Q3), item D-Q2>C-Q2>A-Q2>E-Q2>B-Q2, medium AQ students item D-Q2>E-Q2>C-Q2>A-Q2>B-Q2, and low AQ students on item D-Q2>C-Q2>E-Q2>A-Q2>B-Q2. This means that high, medium, and low AQ students tend to provide evidence most easily in the context of problems related to the manufacture of salt crystals (item B-Q2), most difficult in providing evidence in the context of the item on the use of $CaMgCO_3$ dolomite fertilizer (item D-Q2).

Third, high AQ students have the ability to provide evidence (Q2) on item E-Q3>A-Q3>D-Q3>C-Q3>B-Q3, medium AQ students on item E-Q3>D-Q3>A-Q3>C-Q3>B-Q3, low AQ students on item D-Q3>E-Q3>A-Q3>B-Q3>C-Q3. This means that high and medium AQ students are easiest to provide evidence in the context of problems related to the manufacture of salt crystals (item B-Q3), and low AQ students in the context of the problem of using antacids as magical medicine (item C-Q3). However, it is most difficult to provide justification in the context of the problem of the difference in the acidity level of HCl and H_2SO_4 (item E-Q3) and the use of $CaMgCO_3$ dolomite fertilizer (item D-Q3).

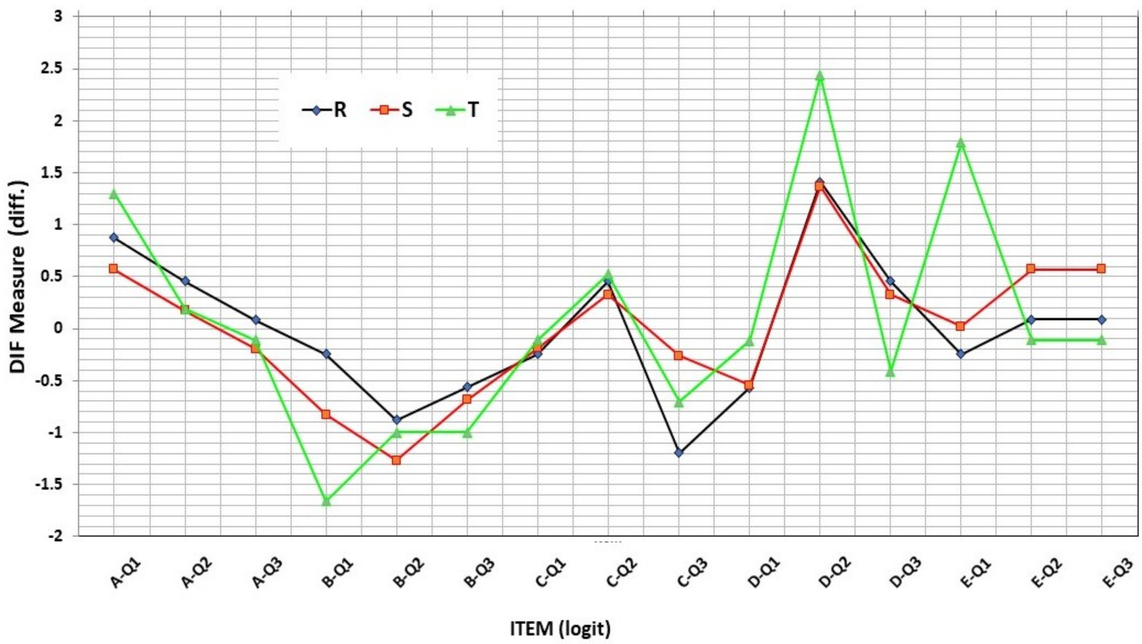

Figure 4. DIF plot of scientific argumentation ability based on differences in adversity quotient ability of students. R = low AQ student criteria, S = medium AQ student criteria, T = high AQ student criteria

Some of the facts/findings above related to the differences in students' scientific argumentation abilities in providing claims (Q1), evidence (Q2), and justification (Q3), on each scientific phenomenon, both based on gender differences, as well as differences in adversity quotient (AQ) abilities, show the weak reasoning abilities of students in connecting the knowledge they have learned with scientific phenomena (Murray et al., 2020). This weak scientific argumentation ability has become a key discussion in science education in the last ten years (Corner & Hahn, 2009; Sadler, 2011). It was also found that students have problems connecting data to support their arguments (Acar et al., 2010). Students who do not understand what a good argument looks like will use intuitive concepts and guesswork in their arguments (Asniar, 2016).

The research findings provide evidence that prospective teachers as the next generation who will conduct face-to-face activities and learning in the classroom, should have good argumentation skills; thus, they can face various problems related to the content and context that will be given to students with various types of children in the future (Asniar, 2016). Student engagement in argumentation requires an explanation or decision on a research question that is supported by some evidence based on empirical data and includes the use of scientific principles, theories, models, and others (Faize et al., 2018).

Then, why do students have difficulty in building their scientific argumentation? It is possible that this is caused by two things. First, it is related to the existence of alternative frameworks or misconceptions (Hoe & Subramaniam, 2016; Laliyo et al., 2022; Suteno et al., 2021), and second, it may be related to students' epistemological experiences in learning. Research reported by Dillon et al. (2006) and Rickinson et al. (2004) explain that even though learning is carried out in laboratory practice or in the field, if students are not given the opportunity to make deeper learning meanings and think critically, then they cannot make cognitive and or affective connections to what they are learning. Therefore, it is a matter of meaningful engagement, in which students are epistemologically involved in contextualizing in forming meaning for their learning. This meaning making aspect of the learning experience is called epistemic engagement (Ryder & Leach, 1999).

The existence of students who have difficulty building scientific arguments about chemical phenomena is thought to occur because misconceptions and aspects of epistemic involvement in learning are not well implemented. According to Kinslow et al. (2018), students' difficulties in forming meaning in their learning are partly due to learning experiences that are less epistemically interesting, where they are involved in data collection practices. Still, they do not think about the function and meaning of the information collected. Therefore, it is important to provide learning tools to integrate complex contextual components, such as interesting issues and themes, into students' learning processes and experiences in order to enhance epistemic engagement and facilitate the improvement of scientific literacy (Astalini et al., 2021; Fadly & Miaturrohmah, 2021; Gulacar et al., 2020; Owens et al., 2019).

Conclusion

The results have shown that this research instrument can be used to measure students' ability to build scientific argumentation. The instrument has good empirical item validity. In addition, it was found that students' abilities were relatively different, especially in terms of making claims, evidence, and justification. This ability is relatively weak, characterized by item sizes that do not fit the construct map. Some items were responded to differently due to gender differences and students' AQ ability. These results suggest that the Rasch model can estimate items that students respond differently to in terms of gender differences and AQ ability and generate valuable information for teachers in developing appropriate and measurable instructional strategies.

Acknowledgments

We would like to thank the Directorate of Research and Community Service (DRPM), Directorate General of Higher Education, and Ministry of Education, Culture, Research and Technology of the Republic of Indonesia for funding this research through the Basic Research Program for Higher Education Excellence at the Research and Community Service Institute of Gorontalo State University, 2022.

References

Acar, O., Turkmen, L., & Roychoudhury, A. 2010. Student difficulties in socio-scientific argumentation and decision-making research findings: Crossing the borders of two research lines. *International Journal of Science Education*, 32(9):1191–1206. https://doi.org/10.1080/09500690902991805.

Asniar, A. 2016. Profil penalaran ilmiah dan kemampuan berargumentasi mahasiswa sains dan non-sains. *Jurnal Penelitian dan Pembelajaran IPA*, 2(1):30-38. https://doi.org/ 10.30870/jppi.v2i1.428.

Astalini, A., Darmaji, D., Kurniawan, D.A., & Chen, D. 2021. Investigating student perceptions based on gender differences using e-module mathematics physics in multiple integral material. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, 9(4):602–619. https://doi.org/10.24815/jpsi.v9i4.21297.

Aydeniz, M., Pabuccu, A., Cetin, P.S., & Kaya, E. 2012. Argumentation and students' conceptual understanding of properties and behaviors of gases. *International Journal of Science and Mathematics Education*, 10(6):1303–1324. https://doi.org/10.1007/ s10763-012-9336-1.

Bond, T.G. & Fox, C.M. 2015. *Applying the Rasch model: Fundamental Measurement in the Human Sciences* (3rd ed.). Routledge Taylor & Francis Group, New York.

Boone, W.J. 2016. Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, 15(4):1–7. https://doi.org/10.1187/cbe.16-04-0148.

Boone, W.J. & Staver, J.R. 2020. Advances in Rasch Analyses in the Human Sciences. *Advances in Rasch Analyses in the Human Sciences*. 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 Dordrecht Heidelberg, New York London. https://doi.org/10.1007/978-94-007-6857-4.

Bradley, J.D. & Mosimege, M.D. 1998. Misconceptions in acids and bases: A comparative study of student teachers with different chemistry backgrounds. *South African Journal of Chemistry*, 51(3):137–143.

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.

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.

Corner, A. & Hahn, U. 2009. Evaluating science arguments: Evidence, uncertainty, and argument strength. *Journal of Experimental Psychology: Applied*, 15(3):199–212. https://doi.org/10.1037/a0016533.

Damanhuri, M.I.M., Treagust, D.F., Won, M., & Chandrasegaran, A.L. 2016. High school students' understanding of acid-base concepts: An ongoing challenge for teachers. *International Journal of Environmental and Science Education*, 11(1):9–27. https://doi.org/10.12973/ijese.2015.284a.

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 uutdoor learning: Evidence from research in the UK and Elsewhere. *School Science Review*, 87:107–111.

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.

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.

Fadly, W. & Miaturrohmah, M. 2021. Arguing skillfully with argument-driven inquiry science textbooks. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, 9(1):119–138. https://doi.org/10.24815/jpsi.v9i1.18630.

Faize, F.A., Husain, W., & Nisar, F. 2018. A critical review of scientific argumentation in science education. *Eurasia Journal of Mathematics, Science and Technology Education*, 14(1):475–483. https://doi.org/10.12973/ejmste/80353.

Gabel, D. 1999. Improving teaching and learning through chemistry education research: A Look to the future. *Journal of Chemical Education*, 76(4):548-553. https://doi.org/10.1021/ed076p548.

Gette, C.R., Kryjevskaia, M., Stetzer, M.R., & Heron, P.R.L. 2018. Probing student reasoning approaches through the lens of dual-process theories: A case study in buoyancy. *Physical Review Physics Education Research*, 14(1):10113. https://doi.org/ 10.1103/PhysRevPhysEducRes.14.010113.

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.

He, P., Liu, X., Zheng, C., & Jia, M. 2016. Using Rasch measurement to validate an instrument for measuring the quality of classroom teaching in secondary chemistry lessons. *Chemistry Education Research and Practice*, 17(2):381–393. https://doi.org/ 10.1039/C6RP00004E.

Herrmann-Abell, C.F. & Deboer, G.E. 2016. Using Rasch modeling and option probability curves to diagnose students' misconceptions. *Paper Presented at the 2016 American Eduacational Research Assossiation Annual Meeting Washington, DC April 8-12, 2016*, 1–12. https://files.eric.ed.gov/fulltext/ED572821.pdf.

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.

Hidayat, W. 2017. Adversity quotient dan penalaran kreatif matematis siswa SMA dalam pembelajaran argument driven inquiry pada materi turunan fungsi. *Kalamatika Jurnal Pendidikan Matematika*, 2(1):15-27. https://doi.org/10.22236/kalamatika.vol2no1. 2017pp15-28.

Hoe, K. & 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.

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, L.A.R., Kilo, A.L., 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 Education*, 21(5):817–835. https://doi.org/10. 33225/jbse/22.21.817.

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. https://doi.org/10.1039/c6rp00137h.

Luxford, C.J. & Bretz, S.L. 2014. Development of the bonding representations inventory to identify student misconceptions about covalent and ionic bonding representations. *Journal of Chemical Education*, 91(3):312–320. https://doi.org/10.1021/ed400700q.

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.

McClary, L.M. & Bretz, S.L. 2012. Development and assessment of a diagnostic tool to identify organic chemistry students' alternative conceptions related to acid strength. *International Journal of Science Education*, 34(15):2317–2341. https://doi.org/10.1080/09500693.2012.684433.

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.

Muhammad, D. 2020. Effect of cooperative learning model type of group investigation and adversity quotient of students on mastery of chemical concepts on salt hydrolysis material. *Jurnal Pendidikan KImia* 12(3):124–135. https://doi.org/10.24114/jpkim.v12i3.21161.

Muntholib, M., Munadhiroh, A., Setiawan, N.C.E., & Yahmin, Y. 2021. High school students' scientific argumentation on chemical equilibrium. *AIP Conference Proceedings*, 2330. https://doi.org/10.1063/5.0043236.

Murray, S.A., Huie, R., Lewis, R., Balicki, S., Clinchot, M., Banks, G., Talanquer, V., & Sevian, H. 2020. Teachers' noticing, interpreting, and acting on students' chemical ideas in written work. *Journal of Chemical Education*. https://doi.org/10.1021/acs.jchemed.9b01198.

Neuman, W.L. 2014. *Social research methods: Qualitative and quantitative approaches (7th ed.).* Pearson Education Limited, United States of America.

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*. 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.

Perera, C.J., Sumintono, B., & Jiang, N. 2018. The psychometric validation of the principal practices questionnaire based on item response theory. *International Online Journal of Educational Leadership*, 2(1):21–38. https://doi.org/10.22452/iojel.vol2no1.3.

Rahayu, S. 2019. Socio-scientific Issues (SSI) in chemistry education: Enhancing both students' chemical literacy & transferable skills. *Journal of Physics: Conference Series*, 1227:012008. https://doi.org/10.1088/1742-6596/1227/1/012008.

Rahayu, S. 2019. Argumentasi ilmiah : Implementasinya dalam pembelajaran kimia untuk meningkatkan keterampilan berkomunikasi. *Prosiding Seminar Nasional Kimia 2019 Jurusan Kimia FMIPA Universitas Negeri Surabaya*, *October 2019*. https://www. researchgate.net/publication/337010572.

Rahayu, S. 2017. *Promoting the 21st century scientific literacy skills through innovative chemistry instruction*. 020025. https://doi.org/10.1063/1.5016018.

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*. Shropshire: Field Studies Council.

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.

Sadler, P.M. 1999. The relevance of multiple-choice testing in assessing science understanding. In J.J. Mintzes, J.H. Wandersee, & J.D. Novak (Eds.), *Assessing science understanding: A human constructivist view* (pp. 251–274). Elsevier Academic Press.

Sadler, T.D. 2011. Socio-scientific issues-based education: What we know about science education in the context of ssi. In T. D. Sadler (Ed.), *Socio-Scientific Issues in the Classroom: Teaching, Learning and Research* (p. Chapter 20, 355–369.). Dordrecht: Springer. https://doi.org/DOI 10.1007/978-94-007-1159-4_16.

Samosa, R.C. 2021. Effectiveness of claim, evidence and reasoning as an innovation to develop students' scientific argumentative writing skills. *Journal of Multidimensional Research & Review*, 2(1):2708–2752.

Sandoval, W.A. 2003. Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences*, 12(1):5–51. https://doi.org/10.1207/ S15327809JLS1201_2.

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.

Shin, N., Choi, S.Y., Stevens, S.Y., & Krajcik, J.S. 2019. The impact of using coherent curriculum on students' understanding of core ideas in chemistry. *International Journal of Science and Mathematics Education*, 17(2):295–315. https://doi.org/10.1007/ s10763-017-9861-z.

Stoltz, P.G. 2018. *Adversity Quotient Mengubah Hambatan Menjadi Peluang*. Jakarta: PT Grasindo.

Sumintono, B. 2014. Model Rasch untuk Penelitian Sosial Kuantitatif. *Makalah Kuliah Umum di Jurusan Statistika, ITS Surabaya, 21 November 2014*, 201:1–9.

Sumintono, B. 2016. Aplikasi Pemodelan Rasch pada Asesmen Pendidikan : Implementasi Penilaian Formatif (assessment for learning). *Makalah Dipresentasikan Dalam Kuliah Umum Pada Jurusan Statistika, Institut Teknologi Sepuluh November, Surabaya, 17 Maret 2016.*, *March*, 1–19.

Sumintono, B. & Widhiarso, W. 2015. *Aplikasi Pemodelan Rasch pada Assessment Pendidikan* (Issue September). Penerbit Trim Komunikata, Cimahi. https://www.researchgate.net/publication/282673464%0AAplikasi.

Suteno, I.K., Laliyo, L.A.R., Iyabu, H., & Abdullah, R. 2021. Mengevaluasi level pemahaman konsep hidrolisis garam peserta didik menggunakan tes diagnostik pilihan ganda empat tingkat. *Jurnal Pendidikan Sains Indonesia (Indonesian Journal of Science Education)*, 9(3):482–497. https://doi.org/10.24815/jpsi.v9i3.20543.

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.

Testa, I., Capasso, G., Colantonio, A., Galano, S., Marzoli, I., Scotti di Uccio, U., Trani, F., & Zappia, A. 2019. Development and validation of a university students' progression in learning quantum mechanics through exploratory factor analysis and Rasch analysis. *International Journal of Science Education*, 41(3):388–417. https://doi.org/10.1080/09500693.2018.1556414.

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.

Tümay, H. 2016. Emergence, learning difficulties, and misconceptions in chemistry undergraduate students' conceptualizations of acid strength. *Science and Education*, 25(1–2):21–46. https://doi.org/10.1007/s11191-015-9799-x.

Tyson, L., Treagust, D.F., & Bucat, R.B. 1999. The complexity of teaching and learning chemical equilibrium. *Journal of Chemical Education*, 76(2–4):554–558. https://doi.org/10.1021/ed077p1560.1.

Ural, E. & Gencoglan, D.M. 2019. The effect of argumentation-based science teaching approach on 8th graders' learning in the subject of acids-bases, their attitudes towards science class and scientific process skills. *Interdisciplinary Journal of Environmental and Science Education*, 16(1):1–15. https://doi.org/10.29333/ijese/6369.

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.