The International Journal of Assessment and Evaluation
ISSN: 2327-7920 (Print), ISSN: 2327-8692 (Online)
Volume 32, Issue 1
https://doi.org/10.18848/2327-7920/CGP/v32i01/49-75

Original Research

Changes in Students' Socioscientific Reasoning in an Environmental Chemistry Class: Application of Multi-Facet Rasch Model

Lukman Abdul Rauf Laliyo, Universitas Negeri Gorontalo, Indonesia
Bambang Sumintono, Universitas Islam Internasional Indonesia, Indonesia
Jafar La Kilo, Universitas Negeri Gorontalo, Indonesia
Citra Panigoro, Universitas Negeri Gorontalo, Indonesia

Received: 05/07/2024; ***Accepted:*** 09/24/2024; ***Published:*** 12/13/2024

Abstract: The primary aim of this research is to evaluate the quality of change in socioscientific reasoning (SSR) among students through the application of the multi-facet Rasch model (MFRM) within a quasi-experimental quantitative research design. The study involved thirty-one students. All participants completed a ten-item test with open-ended questions. Student responses were assessed by five raters using a rubric, and the data were analyzed using the MFRM stacking technique. The analysis revealed that the data aligned with the MFRM measurements, and there were variations in students' SSR scores between T1, T2, and T3. The postintervention change (T1–T2) demonstrated a positive shift, indicating the effectiveness of the intervention. However, postretention changes (T2–T3) showed a negative trend, suggesting a weakening of the intervention's meaningfulness. When examining the nature of postintervention and postretention changes together, it became apparent that a significant proportion of students (71%) exhibited weak and inconsistent changes in their SSR, with some students experiencing anomalous shifts. Importantly for the accuracy of the MFRM measure, although the intervention initially resulted in positive changes, these changes did not persist in the postretention period. This suggests there is room for improvement in the long-term impact of the intervention on students' SSR.

Keywords: *Assessment, Change, Rasch Model, SSI, SSR*

Introduction

One of the serious challenges of science learning in the future is how to create an effective and meaningful learning environment (Fraser 2012; Koul 2023), culminating in positive and productive learning experiences (Rahayu et al. 2021; Rahmawati et al. 2021). As a result, each student is skilled in building their epistemological reasoning framework, commonly known as socioscientific reasoning (SSR) (Sadler et al. 2007). SSR is an important competency, related to students' thinking practices in their efforts to understand and provide solutions to socioscientific issues (SSIs). As SSI is relatively complex, it requires scientific and social considerations, meaning that it can be resolved meaningfully by students (Herman et al. 2021). Recent studies have shown that students' skills in building SSR frameworks are determined by the quality of their epistemological engagement in learning (Kinslow et al. 2018; Kutluca 2021). This view is underpinned by the principle of constructivism learning theory (Bodner 1986; Sesen and Tarhan 2011), that students construct their own epistemic meaning, based on learning experiences, which they obtain through interactions between known, believed facts, data or ideas, and newly discovered ideas (Krahenbuhl 2016; Pande and Bharathi 2020).

Scholars agree that the SSR framework can change and evolve, reflecting changing understandings of how students present and package information (Kinslow et al. 2018). Therefore, epistemological learning experiences tend to determine the development of students' scientific literacy (Sadler and Zeidler 2009). That is why, setting the conditions for SSI-based learning is considerably important and strategic (Herman et al. 2021; Ke et al. 2020), especially if it is placed as a problem context in learning (Kutluca 2021; Zeidler et al. 2013), to navigate the engagement and thought process of students building their epistemological reasoning (Owens et al. 2019). Typically, SSIs are controversial topics and tend to be complicated and complex, but are conceptually bound to science and cannot be solved by a scientific view alone (Corner and Hahn 2009). Such concepts are complex because they involve multiple perspectives and do not allow for a single deduction or solution. For example, issues of pollution and ecological damage, etc. (Irmak 2020; Sadler 2004a, 2004b; Zohar and Nemet 2002).

Previous research has shown that SSI-based learning conditions have been effective in improving the meaningfulness of content knowledge (Sadler et al. 2016), science practices (Peel et al. 2019; Zangori et al. 2017), and the nature of science understanding (Khishfe and Lederman 2006); encouraging the growth of good student personality traits, in their active role as members of society (Lee et al. 2013); and developing higher-order cognitive skills, critical thinking, formal reasoning, decision-making, and argumentation (Sadler and Zeidler 2009; Zeidler et al. 2013). However, some scholars claim that this research has not directly measured the process of students' epistemic engagement with learning complexity and its role in negotiating SSI (Kinslow et al. 2018). Therefore, efforts are needed to increase students' epistemic engagement in SSI-based learning, to a higher level, i.e., the ability to contextualize or make meaning from their own learning experiences (Ryder and Leach 1999). However, these efforts often fail, due to the lack of opportunities and appropriate learning environments for them to engage directly in epistemic activities. For example, many students are active in data collection but are unable to explain what the data they collect means, and how the laboratory tests relate to the problem they are researching (Bonney et al. 2009; Cooper 2012).

Some scholars rely on informal reasoning to interpret students' learning experiences (Sternäng and Lundholm 2012), incorporating sociocultural, economic, and value aspects of the SSI context studied in learning. This method is then applied as an SSR construct (Kinslow et al. 2018; Romine et al. 2017), covering four dimensions of competence, namely: (1) complexity, (2) perspective taking, (3) advanced inquiry, and (4) skepticism (Cian 2019; Hancock et al. 2019; Kinslow et al. 2018). Research related to SSR constructs includes that reported by Romine et al. (2017). This research found that most students' SSR framework did not depend on their declarative knowledge, and there were inconsistent item functions. No significant change in students' SSR was found, due to the influence of the short duration of the SSI-based intervention. Kinslow et al.'s (2018) research found that SSI-based learning tends to support students' SSR development (Sadler et al. 2007, 2011). Research by Owens et al. (2019) found that students need challenging contextual learning to enable them to analyze problems from various perspectives and use their skepticism to analyze potentially biased information. Research by Herman et al. (2021) and Owens et al. (2022) report that learners' SSR is built based on the knowledge domain they understand, in responding to SSI effectively. However, as far as the literature search has been carried out, it has never been tested and revealed to what extent the effectiveness and meaningfulness of learning designed based on SSI, in influencing the quality of SSR, changes. In this study, intervention testing was carried out with different techniques, which emphasized the setting of learning conditions, by integrating SSI as a problem context, with sufficient learning time duration, and using stacking analysis techniques with a multi-facet Rasch model (MFRM) approach.

The authors developed environmental problem-solving–based learning (EPBL), as a research intervention model. This model is adapted from the process oriented guided inquiry learning (POGIL) model (Treagust et al. 2018). In the EPBL model, SSI is integrated as a context for learning problems, so that the learning process can take place factually, through the process of observation, laboratory testing, and environmental exploration (Grooms 2020). The effectiveness and meaningfulness of the intervention were assessed by the quality of change in students' SSR, using a ten-item test with open-ended questions. The test was adapted from the Quantitative Assessment of SocioScientific Reasoning (QuASSR) (Kinslow et al. 2018; Romine et al. 2017; Sadler et al. 2007). The test was administered at three different measurement conditions: preintervention/pretest (T1), postintervention/posttest (T2), and postretention/delay-test (T3). This postretention measurement was conducted after six months of intervention implementation. The purpose is to determine the number of positive SSR changes, which are still able to be maintained, remembered, or applied by students, after six months of intervention. That is, the higher the retention, the more meaningful the intervention carried out in learning (Degeng 1999; Reigeluth 1992).

Students' responses at T1, T2, and T3 were examined by five expert raters and analyzed using the stacking technique with the MFRM approach. This technique has the advantage and uniqueness in its analysis, based on an individual-centered statistical perspective, allowing researchers to estimate carefully and accurately, whether and how each individual student changes during the intervention (Boone and Staver 2020; Laliyo et al. 2022; Ling et al. 2018; Sukmawati 2023; Wright 1996, 2003). MFRM is able to provide a fair and accurate estimation of the size (logit) of each student's SSR change at T1, T2, and T3. In addition, each rater can be defined in terms of their rating, based on the way they rated. This means that MFRM requires that there are no raters whose ratings are identical (Bond and Fox 2015; Boone et al. 2014;

Kudiya et al. 2018). Thus, the estimation of the size (logit) of students' SSR change can be executed independently and simultaneously, allowing the comparison of multiple items at the same time on the same scale (Zhai 2022). At the individual level, differences in T1, T2, and T3 measures can reveal different student responses, specific or not suitable for the intervention (Ling et al. 2018). The research questions are as follows: (1) How good is the suitability of the psychometric attributes of this research instrument when used to measure changes in student SSR? (2) In terms of the acquisition of student SSR measures: (a) Is there a difference in student SSR measures in conditions T1, T2, and T3? (b) What is the nature of students' SSR changes that occurred at postintervention (T1–T2) and at postretention (T2–T3)? (3) If the nature of SSR changes that occurred postintervention and postretention is connected: (a) what is the quality of students' SSR changes? (b) Are there students who experience anomalous SSR changes? If so, are there any other factors that contribute to the anomalous SSR changes?

In conclusion, this article addresses the research problem of assessing changes in students' SSR in an environmental chemistry class, with a focus on evaluating the effectiveness and meaningfulness of a pedagogical intervention. Specifically, the study seeks to understand how students' SSR evolves and the impact of a specific intervention model, EPBL, on this evolution. By utilizing the MFRM, the aim is to investigate the quality of these changes in SSR among students. The central research question guiding this investigation is: "To what extent do changes in student SSR measures occur as a result of an EPBL intervention, and what is the quality and nature of these SSR changes, including any factors contributing to anomalous shifts?"

Methodology

General Background

This study employed a quantitative approach, in a quasi-experiment quantitative research design without a control group (Creswell 2012). The EPBL model intervention is designed based on participatory experiments that present conceptual conflicts, between science and society, as a strategy to encourage students to build SSR frameworks with correct concepts (Almuntasheri et al. 2016). Therefore, students have the opportunity to change their conceptions and epistemological reasoning to be more accurate and scientific (Zvoch et al. 2019). However, it must also be recognized that potentially, there may be students who tend to be resistant to change (Gette et al. 2018).

Participants

The participants in this study involved thirty-one chemistry teacher trainees (aged 20 to 21 years), who were enrolled in an environmental chemistry course at the Chemistry Department, Faculty of Mathematics and Natural Sciences, State University of Gorontalo, Indonesia. All participants, comprising six males and twenty-five females, had passed the basic chemistry course, and the basics of analytical chemistry, as prerequisites before taking environmental chemistry. Traditionally, this course in each regular semester is conducted weekly, with three sections (fifty minutes) of theoretical lectures and one section of laboratory practical lectures, for sixteen weeks. In general, this course is known to be quite “difficult,” as it requires significant face-to-face time, especially in developing problem-solving quantitative information analysis and interpretation. For this reason, some students choose to postpone taking this course until their final year of study. Research permits were obtained through the applicable procedures. The researcher applied for a research permit through the head of the university, and then forwarded it to the head of the study program. After the research plan was approved, perceptions were shared with the teaching team of the environmental chemistry course regarding the schedule and learning strategy. Related to the research code of ethics, students have expressed their willingness, and they were given information about the purpose of the research, procedures, and risks and benefits of being a research subject, and were informed that their identity will be kept confidential (Taber 2014).

Learning Procedure

After completing a pretest (T1), the EPBL model was applied to the environmental chemistry course, for sixteen learning activities, in five learning stages: (1) problem orientation with SSI context, (2) exploration and observation, (3) concept formation and problem identification, (4) application, solution, and follow-up, and (5) reflection and conclusion. The problem orientation stage consists of two activities. Firstly, SSI was introduced as the context of the main learning problem. Students were asked to observe a video presentation and read a research article related to pollution in Lake Limboto that had been provided. Following the observation was asking students to explain what they observed in groups, describe the problem, and formulate the hypotheses. Secondly, students were asked to answer some specific questions, guided by the worksheet. These questions included, for example, is the lake polluted? What is the concentration of dissolved chemical species (e.g., nitrate, nitrite, phosphate, sulfate, dissolved oxygen, pH, etc.) of the Limboto lake water? What is the water quality? What are the possible solutions? These are examples of open-ended, inquisitive questions that cannot be answered in a conventional class.

In the exploration and observation stage, students conducted exploration, observation, and interview activities related to community activities in managing and utilizing the lake. Students went directly to the field, taking water samples at several points around the lake, observing the lake shore, and interviewing the community. The aim is for students to collect important information related to the causes of murky lake water and information about the content of pollutants through laboratory testing. With this flow of activities, students are expected to build a framework for understanding their epistemology, based on evidence from observations, interviews, and laboratory tests. In the concept formation and problem identification stage, students are asked to formulate and categorize problems based on the results of exploration and observation, namely the SSI category that occurs due to the process of cause-and-effect relationships and comprises conflicts between scientific and social aspects. This problem formulation is tested through panel discussions so that communication patterns and thinking constructions can strengthen the meaningfulness of problem-solving.

At the stage of implementing solutions and follow-up, students were asked to present the formulation of solutions and follow-up problem-solving, based on the perspective of stakeholders, and the perspective of students, which were processed from the results of information collection, experiments, and conclusions (Mitarlis et al. 2020; Pedaste et al. 2015). In addition, students were asked to verify and clarify potentially biased information and explain what factors cause the biases. It was expected that at this stage, students' epistemological reasoning developed so they could justify the solution framework and information bias in society. Finally, at the reflection and conclusion stage, each student was asked to compile a conceptual map of the problem based on the learning experience they had participated in. The map must be prepared by including keywords that connect each concept, thus forming a logical and empirical thinking construct. With this concept map, it is expected that the structure of students' thinking in solving the pollution problem in Lake Limboto can be identified.

Instrument and Procedures

The research instrument was adapted from the QuASSR test and developed using a four-stage construct modeling approach (Brown et al. 2010; Hadenfeldt et al. 2013; Wilson 2004). First, the SSR construct was defined. In this study, the construct of SSR is defined as students' thinking practices in understanding and solving the problem of pollution and ecological damage in Lake Limboto, presented in Table 1. Second is item design. The basic structure of this test was designed based on the type of problem description measured in each dimension, which includes each subdimension of SSR. Third is the design of assessment criteria. The scoring criteria are based on a rubric, where each item is given in a score range of 0 to 4, according to the characteristics of the measured subdimension. An example of this is the assessment for item A1. This item measures students' skills in identifying categories of SSI-contextualized problems resulting from pollution and ecological damage in Lake Limboto. A score of 4 is given if students can identify one example of an SSI-contextualized problem and explain it in detail. Score 3 is if students are able to identify one example of a problem with an SSI context, but not explained in detail. Score 2 is if students can identify one example of a problem, but cannot be categorized as a problem with the SSI context. Score 1 is given if one example of a social-scientific context problem category is not identified. Score 0 is given if the student does not provide an answer. This instrument has also been validated in terms of the suitability of the content of each item (subdimension) with the construct (dimension) and the language used. The validation results from five experts were expressed in Fleiss Measure, $K = .96$, meaning that the five validators agreed that the validity of this item was categorized as good.

The fourth is the MFRM analysis approach. The Rasch model was developed by Georg Rasch ([1960] 1980) and has been widely used to assess the feasibility of instruments and diagnose problems (Andrich and Marais 2019), including changes in students' SSR (Kinslow et al. 2018; Romine et al. 2017; Romine and Sadler 2016). The distinctive features of this analytical approach are: (1) converting students' probabilities of item completion into log odds and logits, thereby reducing the problem of nonlinearity in raw scores; (2) placing students' skill level and item difficulty on the same interval scale, so that item scores by particular students can be predicted; and (3) providing information, such as item fit statistics, to check whether each test item used fits the underlying construct, and to identify which students are responding in unexpected ways (Fox and Jones 1998). MFRM was developed by Linacre (1989), which combines measurements of more than one variable or aspect. This approach differs from the typical two-aspect measurement model (Eckes 2015). MFRM allows the analysis of multiple aspects simultaneously and independently at item difficulty levels with the same interval scale (Myford and Wolfe 2003). It also provides a better analysis of how well each facet fits the model and detects potential interaction effects between facets (Bond and Fox 2007; Eckes 2015; Linacre 1989).

Table 1: Conceptual Map of SSR Constructs

Dimensions	*Subdimension / Indicator*	*Item*
A. Identify complex SSI categories	Skillfully identify categories of problems in the SSI context	A1
	Skillfully describe the categories of problems with SSI context, which occur due to the process of cause-and-effect relationships	A2
	Skilled at identifying conflicts between science and sociocultural aspects in complex SSI problem categories	A3
B. Analyzing SSI in various perspectives	Skillfully identify differences in stakeholder opinions	B1
	Skillfully describe differences in stakeholder perspectives	B2
	Skillfully provide solutions for problem-solving	B3
C. Identify aspects of SSI that need further investigation	Skillfully formulate specific research questions and follow-ups	C1
	Skillful in formulating prioritization strategies	C2
D. Categorize potentially biased information with skepticism	Skillfully critically analyze potentially biased information	D1
	Skillfully provide critical arguments as to why information may be biased	D2

The conceptual model of this study, which emphasizes participatory learning through problem-solving in science–society contexts, directly targets the improvement of SSR in each dimension:

1. Dimension A (Identification of complex SSI): The participatory approach encourages students to confront real-world issues, thus promoting a deeper understanding of complex categories of scientific and social problems.
2. Dimension B (Analysis of stakeholder perspectives): By interacting with diverse community perspectives and scientific data, students develop the ability to analyze different viewpoints and propose comprehensive solutions.

3. Dimension C (Further investigation): The fieldwork component of the intervention drives students to formulate investigative questions, encouraging scientific inquiry and application of theoretical knowledge.
4. Dimension D (Skepticism toward biased information): Reflection and data analysis enable students to critically evaluate the reliability of information, fostering skepticism and strengthening their epistemic reasoning.

Hypotheses and Expected Outcomes

The EPBL intervention model, centered around participatory problem-solving of science–society conflicts, is expected to influence students' SSR abilities across four dimensions (as outlined in Table 1). The hypotheses that guide this study align with specific dimensions and subdimensions of SSR. These hypotheses clarify the relationship between the intervention and anticipated changes in students' skills:

1. Hypothesis 1: Students will show significant improvement in their ability to identify complex SSI categories (Dimension A), such as recognizing the cause-and-effect relationships between scientific and sociocultural aspects of environmental problems (A1–A3). The expectation is that after engaging in the EPBL activities, students will become more proficient at distinguishing between science-based and sociocultural perspectives on environmental issues.
    - *Expected Outcome*: Students will move from basic identification of problems (T1) to a more nuanced understanding of conflict categories by the posttest (T3).
2. Hypothesis 2: The intervention will enhance students' skills in analyzing SSI from various perspectives (Dimension B), especially in recognizing stakeholder differences and generating problem-solving strategies (B1–B3). By simulating real-world scenarios and stakeholder roles, students are expected to practice interdisciplinary reasoning and appreciate the diversity of perspectives in addressing environmental problems.
    - *Expected Outcome*: An increase in the depth and clarity of stakeholder analysis and the complexity of proposed solutions between T1, T2, and T3.
3. Hypothesis 3: The intervention will promote students' capacity to identify aspects of SSI requiring further investigation (Dimension C). After exploring pollution-related data and conducting field observations, students are expected to ask more specific research questions and prioritize next steps in the inquiry process (C1–C2).
    - *Expected Outcome*: Students will move from vague or general questions at T1 to more targeted and scientifically grounded research questions and investigation strategies by T3.

4. Hypothesis 4: Students' ability to categorize potentially biased information and apply skepticism (Dimension D) will improve. Through engagement in fieldwork, data analysis, and reflection on biases in media or community knowledge, students are expected to develop stronger critical reasoning skills (D1–D2).
    - *Expected Outcome*: By T3, students should be more adept at recognizing bias in environmental discourse and providing critical arguments, building on their initial responses in T1.

Data Analysis

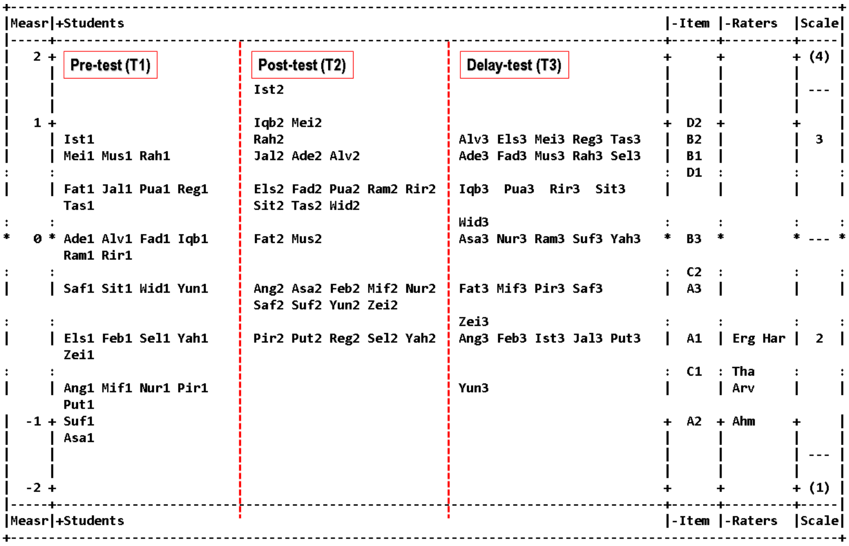
Data collection utilized the same structure of T1, T2, and T3 items. Previously, all students were informed of the data-collection schedule and the mechanism for responding to each item. The test was administered online through the link provided, and students responded in writing in a designated class according to the time allocation provided (120 minutes). Student responses to T1, T2, and T3 were obtained in Excel as description data. This data was then organized by item number according to the T1, T2, and T3 data groups, and assessed by five experts based on the rubric. The scores of the expert assessment results were then arranged using the stacking technique, by placing the T1, T2, and T3 data together vertically (Anselmi et al. 2015; Herrmann-Abell et al. 2013; Ling et al. 2018; Sunjaya et al. 2021; Wright 2003). Each expert rating appeared once in T1, T2, and T3 for each student and item. This data was then converted into data with the same interval scale using Facets Version No. 3.83.3 software (Linacre 2020), thus allowing the researcher to examine changes in each student (Wright 2003). Checks were made on the same items, so it was possible to determine each student's measure of change at T1, T2, and T3 (Wright 2003). Thus, each student produces three measures, i.e., T1, T2, and T3, on each item. The proposed research hypothesis is that students' SSR ability changes from T1 to T2 and T3.

Results

Psychometric Attribute Analysis

Tested the suitability of student, item, and rater psychometric attributes with MFRM, based on the test results of unidimensionality, reliability, person fit, item fit, rater assessment quality, and Wright map. First, unidimensionality. This is one of the key requirements for Rasch measurement (Bond and Fox 2007; Chi et al. 2021; Wang and Willson 2005). Unidimensionality means that the ten items used as measurement instruments in this study measure only one construct at a time, namely SSR skills (Bond and Fox 2007; Boone et al. 2014; Boone and Staver 2020; Fisher 2007; Myford and Wolfe 2004). In this study, unidimensionality was measured using Principal Component Analysis (PCA) of the residuals to determine the extent to which instrument variance measures what it is supposed to measure (Aryadoust et al. 2021; Ding 2018; Tseng et al. 2019). The estimated fit of the data to the model using the Chi-squared test results is 10,370.7715, at the level of significance probability = .0000. This indicates that the data fits the model or the data can be analyzed with MFRM. This overall reference shows that the data fits the model (Boone and Staver 2020). The raw variance explained by measures of this research data is 30.14 percent, greater than the minimum unidimensionality requirement of 20 percent. In other words, the data can be analyzed, because most of the nonrandom variance can be found in the data and is explained by one latent dimension (Chi et al. 2021; Eckes 2015), and the level of item independence in the instrument is very good (Linacre 2020; Ling Lee et al. 2020).

Second, reliability. The Rasch model provides two reliability statistics: the reliability index and the separation index. The separation index measures the spread of an estimate relative to the precision of that measure (Chi et al. 2021). This separation index measure can reach much larger values depending on the size of the error variance (Eckes 2015). These two types of reliability statistics indicate the reliability of measurement differences. According to the test results, the reliability index of assessment items (.99 logits) and students (.88 logits) was very good reliability, because the value was greater than .8 logits, as recommended by Bond and Fox (2007). The acquisition of the stratum index was also higher than 3, which indicates the data collected is reliable (Kudiya et al. 2018). In addition are the student separation index (2.67) and the item separation index (9.84). Both of these index values meet the criteria of 2.0 logits as suggested by Linacre (2020). The magnitude of the student separation index (2.67) and the student reliability index (.88) illustrate the good sensitivity of the instrument, especially in distinguishing high-ability and low-ability students. The acquisition of the item separation index (9.84) and item reliability index (.99) is empirical evidence of the level of student ability and supports the very good construct validity of the instrument (Boone et al. 2014; Boone and Staver 2020; Linacre 2020).



```
+------------------------------------------------------------------------------------------------------------+
|Measr|+Students                                                                      |-Item |-Raters  |Scale|
|-----+-------------------------------------------------------------------------------+------+---------+-----|
|   2 + Pre-test (T1)            Post-test (T2)             Delay-test (T3)           +      +         + (4) |
|     |                                                                              |      |         |     |
|     |                          Ist2                                                |      |         | --- |
|     |                                                                              |      |         |     |
|   1 +                          Iqb2 Mei2                                           +  D2  +         +     |
|     | Ist1                     Rah2                       Alv3 Els3 Mei3 Reg3 Tas3  |  B2  |         |  3  |
|     | Mei1 Mus1 Rah1           Jal2 Ade2 Alv2             Ade3 Fad3 Mus3 Rah3 Sel3  |  B1  |         |     |
:     :                                                                              :  D1  :         :     :
|     | Fat1 Jal1 Pua1 Reg1      Els2 Fad2 Pua2 Ram2 Rir2   Iqb3  Pua3  Rir3  Sit3    |      |         |     |
        Tas1                     Sit2 Tas2 Wid2
:     :                                                    Wid3                      :      :         :     :
*   0 * Ade1 Alv1 Fad1 Iqb1      Fat2 Mus2                  Asa3 Nur3 Ram3 Suf3 Yah3  *  B3  *         * --- *
        Ram1 Rir1
:     :                                                                              :  C2  :         :     :
|     | Saf1 Sit1 Wid1 Yun1      Ang2 Asa2 Feb2 Mif2 Nur2   Fat3 Mif3 Pir3 Saf3       |  A3  |         |     |
                                 Saf2 Suf2 Yun2 Zei2
:     :                                                    Zei3                      :      :         :     :
|     | Els1 Feb1 Sel1 Yah1      Pir2 Put2 Reg2 Sel2 Yah2   Ang3 Feb3 Ist3 Jal3 Put3  |  A1  | Erg Har |  2  |
        Zei1
:     :                                                                              :  C1  : Tha     :     :
|     | Ang1 Mif1 Nur1 Pir1                                 Yun3                      |      | Arv     |     |
        Put1
|  -1 + Suf1                                                                         +  A2  + Ahm     +     |
|     | Asa1                                                                         |      |         |     |
|     |                                                                              |      |         | --- |
|     |                                                                              |      |         |     |
|  -2 +                                                                              +      +         + (1) |
|-----+-------------------------------------------------------------------------------+------+---------+-----|
|Measr|+Students                                                                      |-Item |-Raters  |Scale|
+------------------------------------------------------------------------------------------------------------+
```

Figure 1: Distribution of Student and Item SSR Levels

Third, the Wright map. Figure 1 is a graphic visualization of the Wright map. This map shows how consistent the distribution of student SSR levels, item difficulty levels, and rater quality is. The higher the logit scale, the higher the level of student SSR and the level of item difficulty, and the more difficult the level of scoring by the rater. Conversely, the lower the logit scale, the lower the student SSR level, the lower the item difficulty, and the easier the rater scoring (Boone et al. 2014; Boone and Staver 2020). The first column on the left is a ruler-like scale that ranges from –2 to +2. The second column presents students' SSR skills distribution in conditions T1, T2, and T3 (which fall between –1.0 and +1.0 logits). Only one student in condition T2 (Ist2: 1.42) exceeded all items, and one student in condition T1 (Asa1: –1.29) could not explain all items. The distribution of item difficulty levels in the third column from the left side all occupy locations between –1.0 and +1.0 logits. No items have the same logit size. The fourth column presents the severity/leniency level of the raters. Severity raters, or those who tend to be the most stingy in scoring, are at the top, and lenient raters, or those who tend to be the most generous in scoring, are at the bottom. All distributions of student, item, and rater responses were modeled with zero logit means, meaning that the means of student, item, and rater responses were on the same logit scale.

Table 2: Item Measurement Report

Item	*Measure*	*Model S.E.*	*Outfit*		*PTMEA Correlation*
			MNSQ	*ZSTD*	
A1	–.42	.06	1.06	.8	.31
A2	–1.04	.07	1.18	1.7	.21
A3	–.31	.06	.70	–5.1	.41
B1	.43	.05	1.03	.5	.34
B2	.63	.05	.68	–6.6	.60
B3	–.08	.05	.97	–.4	.52
C1	–.45	.05	1.15	2.0	.34
C2	–.07	.05	1.43	6.4	.48
D1	–.39	.05	.86	–2.6	.47
D2	.92	.05	1.10	1.6	.29

Fourth, item validity (item fit). An item is said to experience misfit if the measurement results obtained do not match three criteria, namely: outfit mean square residual (MNSQ): $.5 < y < 1.5$; outfit standardized mean square residual (ZSTD): $-2 < Z < +2$; and point measure correlation (PTMEA CORR): $.4 < x < .8$ (Boone et al. 2014). Based on the results of item fit testing (Table 2), it is known that there are no items that do not fit the MNSQ outfit criteria. However, four items (A3, B2, C2, D1) do not fit the ZSTD criteria. Five items (A1, A2, C1, C2, D2) did not fit the PTMEA CORR criteria, but none of the items were negative. This evidence shows that none of the items do not meet all three criteria, or no items deviate or experience misfit. This means that all items are appropriate and valid. In addition, it was found that item D2 (.92), was the most difficult item. Critical arguments or reasons why information can be biased; it tends to be the most difficult for students. Item A2 (–1.04) is the easiest item. This means that in terms of the skill of describing categories of SSI-contextualized problems that occur due to the process of cause-and-effect relationships, it tends to be the easiest for students. Three items are higher than the average item size (.00 logit), namely: item D1 (.39), item B1 (.43), and item B2 (.63). The other five items lie below the .00 logit, namely item C1 (–.45), item A1 (–.42), item A3 (–.31), item B3 (–.08), and item C2 (–.07). All ten assessment items have standard errors that are smaller than .15 logits, which means that the measurement precision is very good.

Table 3: Quality of Rater Ratings

Rater	*M*	*Outfit*		*PTMEA Corr.*	*Separation*	*Srata*	*Reliability*	*Chi-Squared*	*Inter-Rater Agreement*	
		MNSQ	*ZSTD*						*Model* (%)	*Expected* (%)
Har	–.38	.84	–4.0	.48	4.79	6.72	.96	.000	55.3	35.6
Tha	–.53	1.01	.10	.49						
Erg	–.58	1.04	1.0	.53						
Arvi	–.79	1.21	4.1	.51						
Ahm	–.92	.98	–.30	.63						

Note: *M* = measure.

Fifth, the quality of raters' ratings. The fourth column in Figure 1 presents information related to the severity level of the raters. Raters who tend to be the most "difficult (stingy) = severe" in giving scores (values) are at the top, while raters who tend to be the most "easy (cheap) = lenient" in giving scores (values) are at the bottom. In Wright's map, it is noted that the distribution of the five raters tends to be very narrow compared to the distribution of students (column 2) and the distribution of assessment items (column 3). This indicates that the five raters work independently in assessing student responses at T1, T2, and T3, and are relatively lenient when viewed from the error on measurement value. The results of testing the quality of raters' assessments (Table 3) show that there are no raters who do not fit the criteria (Boone and Staver 2020; Linacre 2020). All raters performed very well, as indicated by the PTMEA CORR value, which was not negative, and all met the qualifications of good rater quality (Bond and Fox 2015; Boone and Staver 2020). Based on the separation value of 4.79, reliability value of .96, and Chi-squared test result = .000, it shows that the data obtained from five raters have excellent reliability values, and the data can be trusted for further analysis. This is corroborated by the acquisition of an inter-rater agreement value of 55.3 percent, which is not much different from the original value of 35.6 percent, meaning that the five raters work independently. The raters did not cheat each other, and there were no extreme judges. The series of validity and reliability test results that have been conducted indicate that the data collected from the measurement instruments developed in this study are in accordance with Rasch modeling.

Nature of Change in SSR

The nature of change in students' SSR is an important indicator to determine the effectiveness of the intervention, done by comparing the size of each student's logit at T1, T2, and T3 (Table 4). The estimation of the nature of change (positive or negative) is seen from the change in the logit size of students in the postintervention condition (T1–T2) and the condition six months after the intervention or postretention (T2–T3). The nature of change was assessed as positive when the logit measure of SSR postintervention: $T1 < T2$, and postretention: $T2 < T3$. This means students' SSR changed for the better, postintervention and postretention. Conversely, the nature of change is considered negative if in the postintervention condition: $T1 \geq T2$ and in the postretention condition: $T2 \geq T3$. This means that the students' SSR was lower in the postintervention and postretention conditions. Based on Table 4, an interesting fact was that in postintervention, almost all (90.3%) students experienced positive changes; in postretention, less than half (45%) students experienced positive changes. At the individual level, in the postintervention condition, the negative nature of change was identified in three students (6, 9, and 27), and in the retention condition, it was seventeen students (1, 2, 3, 4, 5, 7, 10, 12, 14, 17, 18, 19, 20, 23, 29, 30, 31). This fact shows that the effectiveness of postintervention learning intervention is much better than postretention.

Table 4: Comparison of Student SSR Logit Measures

No	*Student*	*Measure (Logit)*			*Nature of Change in Student SSR*		*Quality of Change*
		T1	*T2*	*T3*	*Postintervention*	*Postretention*	
1	Jal	.24	.56	–.38	Increased/Positive	Declined/Negative	Low
2	Wid	–.14	.36	.24	Increased/Positive	Declined/Negative	Low
3	Sit	–.17	.21	.18	Increased/Positive	Declined/Negative	Low
4	Mei	.43	.94	.78	Increased/Positive	Declined/Negative	Low
5	Rah	.56	.86	.59	Increased/Positive	Declined/Negative	Low
6	Reg	.21	–.56	.82	Declined/Negative	Increased/Positive	*Low
7	Ade	–.01	.46	.38	Increased/Positive	Declined/Negative	Low
8	Fad	–.01	.13	.56	Increased/Positive	Increased/Positive	Good
9	Mus	.59	–.01	.56	Declined/Negative	Increased/Positive	*Low
10	Ist	.74	1.42	–.40	Increased/Positive	Declined/Negative	Low
11	Rir	–.01	.21	.30	Increased/Positive	Increased/Positive	Good
12	Iqb	.04	.94	.27	Increased/Positive	Declined/Negative	Low
13	Nur	–.63	–.17	.07	Increased/Positive	Increased/Positive	Good
14	Ram	.07	.18	.04	Increased/Positive	Declined/Negative	Low
15	Put	–.81	–.56	–.53	Increased/Positive	Increased/Positive	Low
16	Asa	–1.29	–.30	–.09	Increased/Positive	Increased/Positive	Good
17	Fat	.24	–.01	–.35	Increased/Positive	Increased/Positive	*Low
18	Suf	–.91	–.25	–.09	Increased/Positive	Declined/Negative	Low
19	Yun	–.35	–.22	–.87	Increased/Positive	Declined/Negative	Low
20	Ang	–.81	–.27	–.40	Increased/Positive	Declined/Negative	Low
21	Tas	.18	.24	.63	Increased/Positive	Increased/Positive	Good

22	Sel	−.40	−.38	.52	Increased/Positive	Increased/Positive	Good
23	Alv	.02	.52	.63	Increased/Positive	Declined/Negative	Low
24	Els	−.50	.27	.63	Increased/Positive	Increased/Positive	Low
25	Pir	−.74	−.50	−.20	Increased/Positive	Increased/Positive	Good
26	Zei	−.56	−.17	−.17	Increased/Positive	Increased/Positive	Good
27	Saf	−.27	−.33	−.17	Declined/Negative	Increased/Positive	*Low
28	Yah	−.58	−.53	.02	Increased/Positive	Increased/Positive	Good
29	Mif	−.68	−.27	−.33	Increased/Positive	Declined/Negative	Low
30	Feb	−.59	−.13	−.45	Increased/Positive	Declined/Negative	Low
31	Pua	−.36	.21	.13	Increased/Positive	Declined/Negative	Low

Note: * = Anomalous changes.

SSR Quality of Change

The quality of change in SSR reflects the meaningfulness of the learning gains resulting from the intervention. Such a quality is determined by the assessment of the nature of change postintervention, which is connected to postretention, with the following pattern of change. First, the change is considered positive if the nature of change postintervention and postretention is considered positive. Second, the change is rated as low quality if: (1) the nature of postintervention change was positive and postretention was negative, or (2) the nature of postintervention change was negative and postretention was positive, and (3) the nature of postintervention change was negative and postretention was negative. Table 4 shows that only 29 percent or 9 students (8, 11, 13, 16, 21, 22, 25, 26, and 28) experienced changes with good quality. Most of the students (71%) tended to experience low-quality changes, four of which were detected to experience anomalous changes (6, 9, 17, and 27). These anomalous changes were identified based on the negative pattern of SSR changes, the second and third types of changes. That is, in the postintervention condition, students' SSR changed to negative (T2 < T1), but in the postretention condition, it changed to positive (T2 > T3), and or negative (T2 < T3). For example, the anomalous changes experienced by students were 6, 9, 17, and 27.

Relative anomalous changes are not the desired SSR changes. The term "anomalous" describes very distinctive/specific conditions of change that arise outside the influence of the intervention. Figure 2 illustrates a comparison of the size of the T1, T2, and T3 measures of students, who experienced positive, negative, and anomalous changes. Student 8 (Fat) was a positive change, as the item logit measure T1(–.01) < T2(.13) < T3(.56). However, student 1 (Jal) is a negative change, as the item logit measure T1(.24) < T2(.56) > T3(–.38). The change in the logit measure of T3, which is lower than T2 is what causes the quality of change of student 1(Jal) to be negative, but not an anomalous change. This is not the case with the next four students. Student 6 (Reg)'s logit measure: T1(.21) > T2(–.56) < T3(.8), student 9 (Mus): T1(.59) > T2(–.01) < T3(.56), and student 27 (Sal): T1(–.27) > T2(–.33) < T3(–.17) tend to have the same pattern, where the logit measure in condition T1 > T2 and in condition T2 < T3.

The T1 > T2 condition reflects that the intervention does not affect change, while the T2 < T3 condition reflects an increase in the logit size of T2. This increase tends not to align with the achievement of change as a result of learning and is anomalous. A different case of change occurred for student 17 (Fat): T1(.24) > T2(–.01) > T3(–.35). The logit measure in condition T1 > T2, and in condition T2 < T3. This means that 17 (Fat) tended not to change due to the intervention and even became lower than before the intervention. This is also an anomalous fact that deviates from the learning objectives.

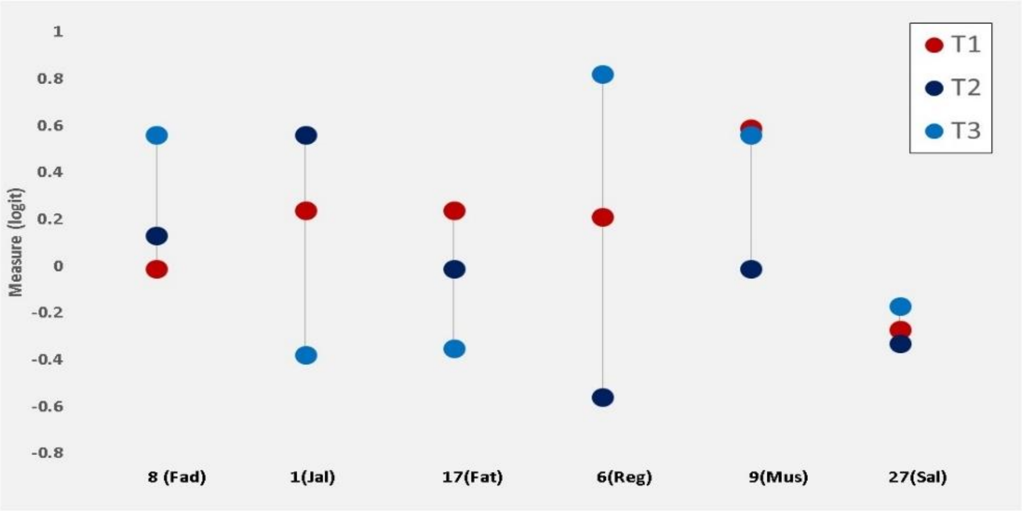


Figure 2: Comparison Graph of SSR Change Patterns of Students Who Are Positive, Negative, and Anomalous

Discussion

Why does the negative and anomalous nature of change occur? This may be due to several factors. First, some students responded differently during the intervention, especially those bored with learning. This psychological factor tends to encourage careless responses and even inhibit their reasoning (Zeidler et al. 2013), making it difficult to build a framework of scientific reasoning and explanation (Rodriguez et al. 2020; Zhai 2022). Second, due to adaptability difficulties, students have difficulty developing their reasoning, and therefore, the frequency of their epistemic engagement in learning is relatively limited (Ryder and Leach 1999). Third, students have difficulty using their declarative science knowledge as a whole (Hancock et al. 2019; Zvoch et al. 2019), thus losing the opportunity to epistemically interpret their learning events (Ryder and Leach 1999). As a result, the level of mastery, accuracy of performance, flexibility of thinking, level of transfer, and retention of learning (Degeng 1999; Reigeluth 1992) tend to weaken, and changes in SSR tend to be volatile and relatively inconsistent.

Fourth, inconsistencies are also caused by students' weak literacy. It is possible that their weak understanding of the content causes them to have difficulty building ideas and problem-solving solutions, or in other words, they have not been able to establish the relevance of content and problem-solving practices (Roberts and Bybee 2014). Despite the SSI-based intervention, students tend to have difficulty developing their scientific knowledge ideas from time to time (Kinslow et al. 2018). Therefore, corrections and refinements to the intervention design are needed, especially in setting conditions that can facilitate students to build epistemic meaning to what they learn (Dillon et al. 2016; Kinslow et al. 2018; Rickinson et al. 2004).

Fifth, inconsistency is also influenced by students' weak knowledge domain. This domain can be in the form of science and nonscience knowledge. Science knowledge is needed to provide general and specific information related to epistemology and scientific literacy, while nonscience knowledge provides SSI information from a sociocultural, moral, and value perspective (Kutluca 2021; Owens et al. 2022; Romine et al. 2017). If the knowledge domain is weak, they may have difficulty understanding the problem being studied and justifying problem-solving (Irmak 2020; Owens et al. 2022). In this research, students' weak functional scientific literacy can be caused by the vulnerability of students' specific knowledge domain related to environmental chemistry content. Therefore, they tend to have difficulty building their SSR.

SSR, as a measure of socioscientific reasoning, depends heavily on students' ability to integrate scientific content knowledge with reasoning skills that address societal issues (Sadler et al. 2007). The anomalies observed in this study may have stemmed from differences in students' baseline knowledge levels, with those possessing a stronger foundation in chemistry, demonstrating more consistent SSR development than their peers. Additionally, external influences not accounted for in the study, such as students' prior exposure to socioscientific discussions or their socio-economic backgrounds, could have impacted their ability to engage with the intervention effectively.

Furthermore, the differences in student responses could reflect varying degrees of cognitive load experienced during the intervention. Students with lower baseline knowledge may have been overwhelmed by the complexity of the socioscientific issues presented, resulting in more erratic or inconsistent reasoning. On the other hand, students with stronger preexisting knowledge may have found the tasks more manageable, leading to more stable and consistent performance. Future studies could benefit from a more nuanced examination of these baseline differences, incorporating pre-tests or more detailed assessments of prior knowledge to better understand how starting points influence SSR outcomes.

Finally, external factors such as classroom dynamics, peer interactions, or even external distractions may have played a role in shaping student responses. These influences, though not controlled for in the study, can create variability in student engagement and cognitive performance, further contributing to the anomalies observed in SSR development. Future research could incorporate more controlled environments or include measures to account for these external influences, providing a clearer understanding of the factors driving inconsistencies in SSR outcomes.

Conclusion

The important findings of this study can show inconsistent changes in students' SSR due to the relatively weak quality of SSR. In addition, it can detail the anomalous changes of students. This is the information that only MFRM-based stacking analysis techniques can provide. Stacking techniques provide information on "who changed" and "what changed" (Barkaoui 2019; Ling et al. 2018; Wright 2003), allowing researchers to detail the nature and quality of change affected by the implementation of an intervention (Laliyo et al. 2022; Pentecost and Barbera 2013; Wright 2003). Although the data generated from the measurement of this research instrument is not rich enough, the potential power of MFRM analysis can explain in detail the changes in SSR down to the student and item level. This is some empirical evidence, as well as showing the superiority and uniqueness of applying the MFRM stacking technique. In addition, this technique can detail how item difficulty changes at T1, T2, and T3 in relation to the content of the skills being measured (Boone and Staver 2020; Pentecost and Barbera 2013; Vorholzer et al. 2020; Zhai 2022), which will be explained in a different article.

In the context of this research, students' relatively weak functional scientific literacy tends to be acceptable for two reasons. First, the science education curriculum has not been designed to be SSI-based. This means that readiness is needed for future curriculum adjustments so that the epistemological framework of reasoning of students in Indonesia can be formed in primary and secondary education. Second, although currently improving scientific literacy has been established as a key competency in the curriculum (Rahayu 2017), it must be complemented by improving students' functional scientific literacy in two scientific visions. Vision I focuses on understanding and developing scientific knowledge in various ways for scientific purposes and taking into account the progress of scientific endeavors. Vision II also uses science concepts and practices but focuses on role and decision-making that encompasses multiple knowledge domains together with science knowledge and informs understanding of and responses to societal issues (Roberts and Bybee 2014). Learning efforts that ignore SSI's integration and the social reasoning dimension will ultimately lead to failure (Owens et al. 2022). In contrast, such efforts will result in a strong argument that Vision II literacy is more functional than Vision I, and SSI serves as a meaningful context for its development (Roberts and Bybee 2014). One means of exploring what students gain, when they engage in socioscientific solution-based learning is SSR (Sadler et al. 2007).

The insights from this study have broad implications for educators across various subjects, not just in environmental chemistry. Teachers can integrate SSI into their curriculum by encouraging students to explore socioscientific issues relevant to their everyday lives, promoting higher-order thinking and decision-making. For instance, in biology lessons, students could investigate ethical debates about genetic engineering or sustainability in agriculture. In physics, teachers could present scenarios related to energy consumption and environmental impacts. Through such approaches, students engage with real-world problems, developing SSR skills in diverse contexts.

Operationalizing these changes at different educational levels requires a strategic approach. At the primary level, SSI-based tasks could involve discussions on simple local issues such as water conservation, with students working together to find solutions, promoting SSR from an early age. At the secondary level, more complex issues such as climate change or renewable energy technologies could be explored, fostering critical analysis and decision-making. Higher education, particularly in STEM fields, could focus on case studies that require students to apply both scientific knowledge and ethical reasoning to global challenges like the COVID-19 pandemic or technological advancements in healthcare.

Institutional constraints, such as varying resources or teacher training, may pose challenges, but solutions are available. In under-resourced schools, for example, online platforms and open-source materials can provide access to case studies and discussions on SSI topics. Furthermore, professional development programs for educators can ensure teachers are equipped to integrate SSI and assess SSR in their classrooms effectively. Workshops and collaborative learning models can be introduced to help teachers develop SSI-based lesson plans that align with the curriculum while enhancing students' SSR abilities.

By applying these strategies, educators across various levels and institutions can strengthen students' functional scientific literacy, ensuring that they are well prepared to tackle complex societal challenges through informed reasoning and decision-making.

Recommendations

The study's limitations prompt several recommendations for future research. Firstly, to address the constraint of small sample size, subsequent studies should prioritize expanding the participant pool. This expansion could involve recruiting students from a broader range of educational institutions across Indonesia to ensure greater diversity and representation within the sample. Similarly, efforts should be made to involve a more varied group of raters in SSR assessment. Diversifying the pool of raters by including individuals from different demographic backgrounds and with varying levels of expertise in chemistry education can enhance the objectivity and comprehensiveness of the evaluation process. Additionally, future research should incorporate measurements of students' mastery of environmental chemistry content to better understand the influence of content knowledge on SSR development. By assessing students' proficiency in this area, researchers can gain insights into the relationship between content mastery and SSR abilities. Furthermore, exploring the interaction effects between demographic differences among raters and students and SSR assessment outcomes can provide valuable insights into potential sources of bias and variation in assessment results. Investigating these interaction effects can inform the refinement of assessment methodologies, ensuring more accurate and equitable evaluations of SSR in future studies.

Limitations

The study faces two primary limitations. Firstly, the sample size of thirty-one students is notably small when considering the entirety of chemistry education students in Indonesia. Additionally, the involvement of only five raters further amplifies concerns regarding the generalizability of the findings. This restricted sample size raises questions about the representativeness of the sample and the breadth of perspectives captured in the assessment process. Secondly, the study's reliance on a limited number of raters poses another limitation. This constraint may introduce bias and hinder the robustness of the assessment process. The lack of diversity among raters, both in terms of demographics and academic backgrounds, could compromise the objectivity and reliability of the SSR evaluation. These limitations highlight the need for future research to address these constraints by involving larger and more diverse samples of both students and raters, thereby enhancing the validity and reliability of the study's findings.

Acknowledgments

The authors 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 Higher Education Excellence Basic Research Program at the Gorontalo State University Research and Community Service Institute, Year 2023.

AI Acknowledgment

The authors acknowledge the use of ChatGPT (https://chat.openai.com) to proofread and edit the writing of this article. The prompts used include "Proofread the article for grammatical accuracy" and "Suggest improvements in clarity and coherence for this section." The output from these prompts was utilized to enhance the article's readability, ensure grammatical correctness, and improve the overall flow of the content.

Informed Consent

The authors have obtained informed consent from all participants.

Conflict of Interest

The authors declare that there is no conflict of interest.

REFERENCES

Almuntasheri, Sultan, Robyn M. Gillies, and Tony Wright. 2016. “The Effectiveness of a Guided Inquiry-Based, Teachers’ Professional Development Programme on Saudi Students’ Understanding of Density.” *Science Education International* 27 (1): 16–39. https://eric.ed.gov/?id=EJ1100181.

Andrich, David, and Ida Marais. 2019. *A Course in Rasch Measurement Theory: Measuring in the Educational, Social, and Health Sciences*. Springer.

Anselmi, Patrizia, Gianmarco Vidotto, Olivia Bettinardi, and Giuseppe Bertolotti. 2015. “Measurement of Change in Health Status with Rasch Models.” *Health and Quality of Life Outcomes* 13 (1): 1–7. https://doi.org/10.1186/s12955-014-0197-x.

Aryadoust, Vahid, Lee Y. Ng, and Hiroki Sayama. 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.

Barkaoui, Khaled. 2019. “Multifaceted Rasch Analysis for Test Evaluation.” In *Companion to Language Assessment Multifaceted*, edited by Khaled Barkaoui. Wiley.

Bodner, George M. 1986. “Constructivism: A Theory of Knowledge.” *Journal of Chemical Education* 63 (10): 873–878. https://doi.org/10.1021/ed063p873.

Bond, Trevor G., and Christine M. Fox. 2007. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 2nd ed. Routledge.

Bond, Trevor G., and Christine M. Fox. 2015. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. 3rd ed. Routledge.

Bonney, Rick, Heidi Ballard, Rebecca Jordan, et al. 2009. *Public Participation in Scientific Research: Defining the Field and Assessing Its Potential for Informal Science Education*. A CAISE Inquiry Group Report, July. https://eric.ed.gov/?id=ED519688.

Boone, William J., and John R. Staver. 2020. “Correction to: Advances in Rasch Analyses in the Human Sciences.” In *Advances in Rasch Analyses in the Human Sciences*, edited by William J. Boone and John R. Staver. Springer.

Boone, William J., Mary S. Yale, and John R. Staver. 2014. *Rasch Analysis in the Human Sciences*. Springer.

Brown, Nathaniel J. S., Erin M. Furtak, Margaret Timms, Shizuka O. Nagashima, and Mark Wilson. 2010. “The Evidence-Based Reasoning Framework: Assessing Scientific Reasoning.” *Educational Assessment* 15 (3): 123–141. https://doi.org/10.1080/10627197.2010.530551.

Chi, Shaohui, Xiufeng Liu, and Zuhao Wang. 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.

Cian, Heidi. 2019. "Influence of Student Values, Knowledge, and Experience on Measures of High-School Student Socio-Scientific Reasoning." PhD diss., Clemson University.

Cooper, Caren B. 2012. "Links and Distinctions Among Citizenship, Science, and Citizen Science." *Democracy and Education* 20 (2): 1–4. https://democracyeducationjournal.org/home/vol20/iss2/13/.

Corner, Adam, and Ulrike Hahn. 2009. "Evaluating Science Arguments: Evidence, Uncertainty, and Argument Strength." *Journal of Experimental Psychology: Applied* 15 (3): 199–212. https://doi.org/10.1037/a0016533.

Creswell, John W. 2012. *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. 4th ed. Pearson Education.

Degeng, I. Nyoman S. 1999. *Teaching Science Taxonomy of Variables*. Ministry of Education and Culture, Directorate General of Higher Education, Project for the Development of Educational Institutions for Educational Personnel.

Dillon, Justin, Mark Rickinson, and Katie Teamey. 2016. "The Value of Outdoor Learning: Evidence from Research in the UK and Elsewhere." In *Towards a Convergence Between Science and Environmental Education.* Routledge.

Ding, Lin. 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.

Eckes, Thomas. 2015. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. 2nd ed. Peter Lang.

Fisher, William P. 2007. "Rating Scale Instrument Quality Criteria." *Rasch Measurement Transactions* 21 (1): 1095. https://www.rasch.org/rmt/rmt211m.htm.

Fox, Christine M., and James A. Jones. 1998. "Uses of Rasch Modeling in Counseling Psychology Research." *Journal of Counseling Psychology* 45 (1): 30–45. https://psycnet.apa.org/buy/1997-42745-003.

Fraser, Barry J. 2012. "Classroom Learning Environments: Retrospect, Context, and Prospect." In *Second International Handbook of Science Education*, vol. 1, edited by Barry J. Fraser, Kenneth G. Tobin, and Campbell J. McRobbie. Springer.

Gette, Colin R., Margarita Kryjevskaia, Matthew R. Stetzer, and Paula R. L. Heron. 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.

Grooms, Jeffery. 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 (8): 2057–2064. https://doi.org/10.1021/acs.jchemed.0c00026.

Hadenfeldt, Jens-Christoph, Sebastian Bernholt, Xiufeng Liu, Knut Neumann, and Ilka Parchmann. 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.

Hancock, Tamara S., Patricia J. Friedrichsen, Allison T. Kinslow, and Troy D. Sadler. 2019. “Selecting Socio-Scientific Issues for Teaching: A Grounded Theory Study of How Science Teachers Collaboratively Design SSI-Based Curricula.” *Science & Education* 28 (6–7): 639–667. https://doi.org/10.1007/s11191-019-00065-x.

Herman, Benjamin C., Marsha H. Newton, and Dana L. Zeidler. 2021. “Impact of Place-Based Socioscientific Issues Instruction on Students’ Contextualization of Socioscientific Orientations.” *Science Education* 105 (4): 585–627. https://doi.org/10.1002/sce.21618.

Herrmann-Abell, Cari F., Jennifer C. Flanagan, and Joseph E. Roseman. 2013. “Developing and Evaluating an Eighth Grade Curriculum Unit That Links Foundational Chemistry to Biological Growth: Using Student Measures to Evaluate the Promise of the Intervention.” Paper presented at the 2013 NARST Annual International Conference, Rio Grande, Puerto Rico, April 6–9, 2013.

Irmak, Mustafa. 2020. “Socioscientific Reasoning Competencies and Nature of Science Conceptions of Undergraduate Students from Different Faculties.” *Science Education International* 31 (1): 65–73. https://doi.org/10.33828/sei.v31.i1.7.

Ke, Li, Troy D. Sadler, Lisa Zangori, and Patricia J. Friedrichsen. 2020. “Students’ Perceptions of Socio-Scientific Issue-Based Learning and Their Appropriation of Epistemic Tools for Systems Thinking.” *International Journal of Science Education* 42 (8): 1339–1361. https://doi.org/10.1080/09500693.2020.1759843.

Khishfe, Rola, and Norman Lederman. 2006. “Teaching Nature of Science Within a Controversial Topic: Integrated Versus Nonintegrated.” *Journal of Research in Science Teaching* 43 (4): 395–418. https://doi.org/10.1002/tea.20137.

Kinslow, Amanda T., Troy D. Sadler, and Hoa T. Nguyen. 2018. “Socio-Scientific Reasoning and Environmental Literacy in a Field-Based Ecology Class.” *Environmental Education Research* 25 (3): 388–410. https://doi.org/10.1080/13504622.2018.1442418.

Koul, Ravinder B. 2023. “Classroom Learning Environments and Assessment Practices in Science Classrooms in Western Australia.” In *Effective Teaching Around the World*, vol. I, edited by R. Maulana, M. Helms-Lorenz, and R. M. Klassen. Springer.

Krahenbuhl, Kevin S. 2016. “Student-Centered Education and Constructivism: Challenges, Concerns, and Clarity for Teachers.” *Clearing House: A Journal of Educational Strategies, Issues and Ideas* 89 (3): 97–105. https://doi.org/10.1080/00098655.2016.1191311.

Kudiya, Kusnadi, Bambang Sumintono, and Sabar Sabana. 2018. “Batik Artisans’ Judgment of Batik Wax Quality and Its Criteria: An Application of the Many-Facets Rasch Model.” In *Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings*, edited by Quan Zhang. Springer.

Kutluca, Arzu Yasemin. 2021. “An Investigation of Elementary Teachers’ Pedagogical Content Knowledge for Socioscientific Argumentation: The Effect of a Learning and Teaching Experience.” *Science Education* 105 (4): 743–775. https://doi.org/10.1002/sce.21624.

Laliyo, Lusdin A., Bambang Sumintono, and Cholifah Panigoro. 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.

Lee, Hyunju, Jinwoong Yoo, Kwangsu Choi, et al. 2013. “Socioscientific Issues as a Vehicle for Promoting Character and Values for Global Citizens.” *International Journal of Science Education* 35 (12): 2079–2113. https://doi.org/10.1080/09500693.2012.749546.

Linacre, John M. 1989. *Many-Faceted Rasch Measurement*. MESA Press.

Linacre, John M. 2020. *A User’s Guide to WINSTEPS® MINISTEP Rasch-Model Computer Programs: Program Manual 4.5.1*. https://cir.nii.ac.jp/crid/1370567187572937223.

Ling Lee, Woei, Karuthan Chinna, and Bambang Sumintono. 2020. “Psychometrics Assessment of HeartQoL Questionnaire: A Rasch Analysis.” *European Journal of Preventive Cardiology* 28 (12): 1–6. https://doi.org/10.1177/2047487320902322.

Ling, Muhee, Vincent Pang, and Cecilia C. Ompok. 2018. “Measuring Change in Early Mathematics Ability of Children Who Learn Using Games: Stacked Analysis in Rasch Measurement.” In *Pacific Rim Objective Measurement Symposium (PROMS) 2016 Conference Proceedings*, edited by Quan Zhang. Springer.

Mitarlis, Mitarlis, Suripto Ibnu, Sri Rahayu, and Sutrisno Sutrisno. 2020. “The Effectiveness of New Inquiry-Based Learning (NIBL) for Improving Multiple Higher-Order Thinking Skills (M-HOTS) of Prospective Chemistry Teachers.” *European Journal of Educational Research* 9 (3): 1309–1325. https://doi.org/10.12973/eu-jer.9.3.1309.

Myford, Carol M., and Edward W. Wolfe. 2003. “Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I.” *Journal of Applied Measurement* 4 (4): 386–422.

Myford, Carol M., and Edward W. Wolfe. 2004. “Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II.” *Journal of Applied Measurement* 5 (2): 189–227.

Owens, David C., Troy D. Sadler, and Patricia J. Friedrichsen. 2019. “Teaching Practices for Enactment of Socio-Scientific Issues Instruction: An Instrumental Case Study of an Experienced Biology Teacher.” *Research in Science Education* 51: 375–398. https://doi.org/10.1007/s11165-018-9799-3.

Owens, David C., Troy D. Sadler, Daniel N. Petitt, and Cory T. Forbes. 2022. "Exploring Undergraduates' Breadth of Socio-Scientific Reasoning Through Domains of Knowledge." *Research in Science Education* 52 (6): 1643–1658. https://doi.org/10.1007/s11165-021-10014-w.

Pande, Madhu, and S. Varun Bharathi. 2020. "Theoretical Foundations of Design Thinking—A Constructivism Learning Approach to Design Thinking." *Thinking Skills and Creativity* 36: 100637. https://doi.org/10.1016/j.tsc.2020.100637.

Pedaste, Margus, Marika Mäeots, Leo A. Siiman, et al. 2015. "Phases of Inquiry-Based Learning: Definitions and the Inquiry Cycle." *Educational Research Review* 14: 47–61. https://doi.org/10.1016/j.edurev.2015.02.003.

Peel, Amanda, Lisa Zangori, Patricia Friedrichsen, Emily Hayes, and Troy Sadler. 2019. "Students' Model-Based Explanations About Natural Selection and Antibiotic Resistance Through Socio-Scientific Issues-Based Learning." *International Journal of Science Education* 41 (4): 510–532. https://doi.org/10.1080/09500693.2018.1564084.

Pentecost, Tom C., and Joseph Barbera. 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, Sri. 2017. "Promoting the 21st Century Scientific Literacy Skills Through Innovative Chemistry Instruction." *AIP Conference Proceedings* 1911 (1): 020025-1–020025-8. https://doi.org/10.1063/1.5016018.

Rahayu, Wahyu, M. D. K. Putra, Y. Rahmawati, B. Hayat, and Ravinder B. Koul. 2021. "Validating an Indonesian Version of the What Is Happening in This Class? (WIHIC) Questionnaire Using a Multidimensional Rasch Model." *International Journal of Instruction* 14 (2): 919–934. https://doi.org/10.29333/iji.2021.14252a.

Rahmawati, Yuli, Elizabeth Taylor, Peter Charles Taylor, and Ravinder Koul. 2021. "Student Empowerment in a Constructivist Values Learning Environment for a Healthy and Sustainable World." *Learning Environments Research* 24 (3): 451–468. https://doi.org/10.1007/s10984-020-09336-9.

Rasch, Georg. (1960) 1980. *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press/Danmarks Paedogogiske Institut.

Reigeluth, Charles M. 1992. "Elaborating the Elaboration Theory." *Educational Technology Research and Development* 40 (3): 80–86. https://doi.org/10.1007/BF02296844.

Rickinson, Mark, Justin Dillon, Kelly Teamey, et al. 2004. *A Review of Research on Outdoor Learning*. National Foundation for Educational Research and King's College London, Field Studies Council.

Roberts, Douglas A., and Rodger W. Bybee. 2014. "Scientific Literacy, Science Literacy, and Science Education." In *Handbook of Research in Science Education*, vol. 2, edited by Sandra K. Abell and Norman G. Lederman. Routledge.

Rodriguez, Jennifer G., Katherine H. Hunter, Lauren J. Scharlott, and Nathaniel M. Becker. 2020. "A Review of Research on Process Oriented Guided Inquiry Learning: Implications for Research and Practice." *Journal of Chemical Education* 97 (10): 3506–3520. https://doi.org/10.1021/acs.jchemed.0c00355.

Romine, William L., and Troy D. Sadler. 2016. "Measuring Changes in Interest in Science and Technology at the College Level in Response to Two Instructional Interventions." *Research in Science Education* 46 (3): 309–327. https://doi.org/10.1007/s11165-014-9452-8.

Romine, William L., Troy D. Sadler, and Angela T. Kinslow. 2017. "Assessment of Scientific Literacy: Development and Validation of the Quantitative Assessment of Socio-Scientific Reasoning (QuASSR)." *Journal of Research in Science Teaching* 54 (2): 274–295. https://doi.org/10.1002/tea.21368.

Ryder, Jim, and John Leach. 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, Troy D. 2004a. "Informal Reasoning Regarding Socioscientific Issues: A Critical Review of Research." *Journal of Research in Science Teaching* 41 (5): 513–536. https://doi.org/10.1002/tea.20009.

Sadler, Troy D. 2004b. "Moral and Ethical Dimensions of Socioscientific Decision-Making as Integral Components of Scientific Literacy." *Science Education* 13 (1): 39–48. https://eric.ed.gov/?id=ED481210.

Sadler, Troy D., and Dana L. Zeidler. 2009. "Scientific Literacy, PISA, and Socioscientific Discourse: Assessment for Progressive Aims of Science Education." *Journal of Research in Science Teaching* 46 (8): 909–921. https://doi.org/10.1002/tea.20327.

Sadler, Troy D., Mary L. Klosterman, and Mustafa Serdar Topcu. 2011. "Learning Science Content and Socioscientific Reasoning Through Classroom Explorations of Global Climate Change." In *Socioscientific Issues in the Classroom: Teaching, Learning and Research*, edited by Troy D. Sadler. Springer.

Sadler, Troy D., Sasha A. Barab, and Brianna Scott. 2007. "What Do Students Gain by Engaging in Socioscientific Inquiry?" *Research in Science Education* 37 (4): 371–391. https://doi.org/10.1007/s11165-006-9030-9.

Sadler, Troy D., William L. Romine, and Mustafa Serdar Topçu. 2016. "Learning Science Content Through Socio-Scientific Issues-Based Instruction: A Multi-Level Assessment Study." *International Journal of Science Education* 38 (10): 1622–1635. https://doi.org/10.1080/09500693.2016.1204481.

Sesen, Burcin Ates, and Leman Tarhan. 2011. "Active-Learning Versus Teacher-Centered Instruction for Learning Acids and Bases." *Research in Science & Technological Education* 29 (2): 205–226. https://doi.org/10.1080/02635143.2011.581630.

Sternäng, Lotten, and Cecilia Lundholm. 2012. “Climate Change and Costs: Investigating Students’ Reasoning on Nature and Economic Development.” *Environmental Education Research* 18 (3): 417–436. https://doi.org/10.1080/13504622.2011.630532.

Sukmawati, Widianingsih. 2023. “Analysis of Changes in Students’ Scientific Literacy Ability After Attending Lectures Using the RADEC Model.” *Jurnal Penelitian Pendidikan IPA* 9 (3): 1039–1044.

Sunjaya, Dian Kusuma, Destri Melina Dewi Herawati, Nanda Indraswari, Gabriella Megawati, and Bambang Sumintono. 2021. “Training and Assessing Model for the Ability of Community Health Volunteers in Anthropometric Measurement Using the Rasch Stacking and Racking Analyses.” *Journal of Environmental and Public Health* 2021: 1–11. https://doi.org/10.1155/2021/5515712.

Taber, Keith Stephen. 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.

Treagust, David F., Shamim S. Qureshi, Velliaris R. Vishnumolakala, et al. 2018. “Process-Oriented Guided Inquiry Learning (POGIL) as a Culturally Relevant Pedagogy (CRP) in Qatar: A Perspective from Grade 10 Chemistry Classes.” *Research in Science Education* 50: 813–831. https://doi.org/10.1007/s11165-018-9712-0.

Tseng, Wen-Ta, Tung-Yu Su, and John Michael Laurence Nix. 2019. “Validating Translation Test Items via the Many-Facet Rasch Model.” *Psychological Reports* 122 (2): 748–772. https://doi.org/10.1177/0033294118768664.

Vorholzer, Andreas, Claudia von Aufschnaiter, and William J. Boone. 2020. “Fostering Upper Secondary Students’ Ability to Engage in Practices of Scientific Investigation: A Comparative Analysis of an Explicit and an Implicit Instructional Approach.” *Research in Science Education* 50 (1): 333–359. https://doi.org/10.1007/s11165-018-9691-1.

Wang, Wen-Chung, and Mark Willson. 2005. “Exploring Local Item Dependence Using a Random-Effects Facet Model.” *Applied Psychological Measurement* 29 (4): 296–318. https://doi.org/10.1177/0146621605276281.

Wilson, Mark. 2004. *Constructing Measures: An Item Response Modeling Approach*. 1st ed. Routledge.

Wright, Benjamin D. 1996. “Time 1 to Time 2 (Pre-Test to Post-Test) Comparison: Racking and Stacking.” *Rasch Measurement Transactions* 10 (1): 478.

Wright, Benjamin D. 2003. “Rack and Stack: Time 1 vs. Time 2 or Pre-Test vs. Post-Test.” *Rasch Measurement Transactions* 17 (1): 905–906. https://www.rasch.org/rmt/rmt171a.htm.

Zangori, Laura, Andrea Peel, Angela Kinslow, Patricia Friedrichsen, and Troy D. Sadler. 2017. “Student Development of Model-Based Reasoning About Carbon Cycling and Climate Change in a Socio-Scientific Issues Unit.” *Journal of Research in Science Teaching* 54 (10): 1249–1273. https://doi.org/10.1002/tea.21404.

Zeidler, Dana L., Benjamin C. Herman, Michael Ruzek, Aaron Linder, and Shu Shan Lin. 2013. "Cross-Cultural Epistemological Orientations to Socioscientific Issues." *Journal of Research in Science Teaching* 50 (3): 251–283. https://doi.org/10.1002/tea.21077.

Zhai, Xiaoming. 2022. "Assessing High-School Students' Modeling Performance on Newtonian Mechanics." *Journal of Research in Science Teaching* 59 (8): 1313–1353. https://doi.org/10.1002/tea.21758.

Zohar, Anat, and Flora Nemet. 2002. "Fostering Students' Knowledge and Argumentation Skills Through Dilemmas in Human Genetics." *Journal of Research in Science Teaching* 39 (1): 35–62. https://doi.org/10.1002/tea.10008.

Zvoch, Keith, Shannon Holveck, and Lori Porter. 2019. "Teaching for Conceptual Change in a Density Unit Provided to Seventh Graders: A Comparison of Teacher- and Student-Centered Approaches." *Research in Science Education* 51: 1395–1421. https://doi.org/10.1007/s11165-019-09907-8.

ABOUT THE AUTHORS

Lukman Abdul Rauf Laliyo: Professor, Department of Chemistry, Universitas Negeri Gorontalo, Gorontalo, Indonesia
Corresponding Author's Email: lukman.laliyo019@gmail.com

Bambang Sumintono: Head of Doctoral Program in Education, Faculty of Education, Universitas Islam Internasional Indonesia, Depok, Indonesia
Email: bambang.sumintono@uiii.ac.id

Jafar La Kilo: Assistant Professor, Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Negeri Gorontalo, Gorontalo, Indonesia
Email: jafar.chem@ung.ac.id

Citra Panigoro: Assistant Professor, Department of Management of Aquatic Resources, Faculty of Fisheries and Marine Technology, Universitas Negeri Gorontalo, Gorontalo, Indonesia
Email: citrapanigoro@ung.ac.id