



**Abstract.** *Scientific reasoning ability is essential to get developed in the current digital age, particularly in the process of judgement and decision-making in complex problems. Few studies have conducted an in-depth exploration of scientific reasoning ability, especially in relation to the confidence level and gender. The scientific reasoning ability of Indonesian upper-secondary school and university students were examined and compared with previous recorded data of US students. In this study, the data were collected from 372 university and 528 upper-secondary education students in Indonesia. Students' scientific reasoning ability was measured using a scientific formal reasoning test (FORT). In addition, confidence level and metacognitive data was collected through self-reported measures. Two-way ANOVA was performed to compare mean differences between groups based on academic level and gender and to observe interaction between the variables. Students' confidence level in selecting the correct answer and distractor answer was analyzed using an independent t-test. The results reveal that many Indonesian students selected specific distractors with relatively high confidence. Moreover, upper-secondary school students and female students selected more distractors than the groups' counterparts. Finally, the factors related to Indonesian students' responses to the scientific formal reasoning were discussed.*

**Keywords:** *confidence level, distractor analysis, gender differences, scientific (formal) reasoning test, scientific reasoning ability, Indonesian student*

**Minsu Ha, Yustika Sya'bandari,  
Ai Nurlaelasari Rusmana**  
Kangwon National University, South Korea  
**Rahmi Qurota Aini**  
Middle Tennessee State University, United States  
**Sarah Meilani Fadillah**  
Kangwon National University, South Korea



## COMPREHENSIVE ANALYSIS OF THE FORT INSTRUMENT: USING DISTRACTOR ANALYSIS TO EXPLORE STUDENTS' SCIENTIFIC REASONING BASED ON ACADEMIC LEVEL AND GENDER DIFFERENCE

**Minsu Ha,  
Yustika Sya'bandari,  
Ai Nurlaelasari Rusmana,  
Rahmi Qurota Aini,  
Sarah Meilani Fadillah**

### Introduction

Science education is crucial for students because science is used to make sense of everyday life, thus the National Research Council (2012), along with several scholars, have advocated scientific reasoning ability for students. For example, research has examined the factors affecting students' understanding of scientific concepts, which also affects their scientific reasoning (Allchin & Zemplén, 2020; Saad et al., 2017; Wilhelm et al., 2018; Yang et al., 2002). Some studies have explored students' scientific reasoning and conceptual change, with relation to course achievement as well (Lawson et al., 2007; She & Liao, 2010; Weld et al., 2011). In addition, Čavojevová and colleagues (2020) have noted that scientific reasoning ability promotes decision-making skill which is aligned with the increased use of technology in recent years (Duan, et al., 2019). Therefore, the goal of science education is to develop scientific reasoning ability in students and help them to engage with science in daily contexts (Kind & Osborne, 2017; van der Graaf et al., 2019).

Several studies by Kuhn et al. (1988, 1989, 1992, and 2004) have extensively contributed to the understanding of scientific reasoning. According to Kuhn, scientific reasoning is focused on conceptualizing individual tasks. Piaget's theory of cognitive development (1965) includes the ability to use abstract thought and logical reasoning, and draw conclusions based on the type of formal operational reasoning for adolescents. Further, Lawson (1978 and 2000) and Kalinowski and Willoughby (2019) have defined scientific reasoning as being identical with formal operational reasoning. These scientific reasoning views refer to the scientific thinking skills for identifying and deriving conclusions from natural processes, including resolving the problem about hypothesis testing, correlation, proportion, probability, and control of variables (COV).

### Research Problem

Global data about students' achievements in science, particularly focusing on scientific competencies such as interpreting data and evidence from scientific perspective are presented in the Programme for International Student Assessment (PISA). Based on the PISA result in 2018, Indonesia, as the fourth populous country right below the US, derived lower student performance in all subjects (reading, mathematics, and science), while students in the US derived average scores in reading, lower average scores in mathematics, and higher average scores in science (OECD, 2018). The score difference between the top 10% and low 10% for Indonesian students in science is the lowest among countries that participated in PISA. Both performances of students in science have remained the same for the overall period. Although several studies have examined scientific reasoning ability for students in the US (Bao et al., 2009; Jensen et al., 2017; Lawson et al., 2007), few have explored it among Indonesian students.

The use of proper assessment is a key factor for measuring students' scientific reasoning ability and it helps to uncover students' perspective of the scientific process. In recent years, Lawson's Classroom Test of Scientific Reasoning (CTSR) (1978, 2000) and the Montana State University Formal Reasoning Test (MSU-FORT) (2019), which are multiple-choice scientific reasoning tests, have been extensively used to measure knowledge. Multiple-choice questions consist of a problem stem or prompt (i.e., the question) and several options that represent potential answers (Tozoglu et al., 2004). Multiple-choice questions have several advantages. Besides it can be scored quickly, it is also cheap, flexible, and free from ambiguity; and the difficulty level can be controlled (Warren, 1979). Recently, the use of multiple-choice questions in tests has become increasingly popular due to technological inventions that aid in analyzing students' response patterns and detecting students' alternative conception (Douglas et al., 2012; Mingo et al., 2018).

### Literature Review

#### *Measuring Scientific Reasoning*

Understanding "scientific reasoning" begins with observing the cognitive development stage noted by Piaget. Piaget stated that cognitive development is divided into four major stages: sensory motor stage, preoperational stage, concrete operational stage, and formal operational stage. Among these stages of thinking, the formal operational stage, to which hypothesis-deductive thinking can be applied, has been considered the highest cognitive development stage (Inhelder et al., 1958). Lawson (1978) defined this formal operational stage as a stage of scientific reasoning wherein hypothesis reasoning is possible, including isolation and COV, combinatorial reasoning, correlational reasoning, probabilistic reasoning, and proportional reasoning. Hawkins and Pea (1987) also perceived scientific reasoning as a concept similar to *critical thinking* (Han, 2013), which can be defined as *hypothesis-deductive thinking* and refers to the development of a hypothesis to answer causal questions by using accurate observations of natural phenomena and several processes to test them (Chamberlin, 1898; Lawson & Lawson, 1980; Platt, 1964).

According to Burmester (1952) and Lawson (1995), scientific reasoning ability is a key component of creativity and critical thinking. Therefore, several studies have attempted to develop measuring tools for students' scientific reasoning ability. In the 1960s and 1970s, many researchers aimed to develop test tools to identify students' reasoning abilities and recognized that the most accurate and informative tool was a clinical interview. However, this tool took a considerable amount of time, sometimes requiring specific equipment to proceed, and most importantly, was difficult to conduct and analyze (Han, 2013). Therefore, scholars aimed to derive a method that can be scored objectively, and that any trained researcher (or teacher) can conduct. As a result, most of the tools were developed using paper and pencil (Burney, 1974; Han, 2013; Longeot, 1965; Raven, 1973; Tobin & Capie, 1980). This part reviewed some tools that were developed for measuring students' scientific reasoning skills and identified their advantages and disadvantages.

First, Lawson (1978), a leading scholar in scientific reasoning studies, developed CTSR as a measuring tool for diagnosing students' scientific reasoning ability. Lawson noted that the paper and pencil method tended to test reading and writing ability, rather than formal reasoning, and he argued that the methods used by Shayer and Wharry (1975), who developed interview methods that minimized time and equipment requirements, were insufficiently diverse. Therefore, Lawson developed and researched items for various measurements, including the following five dimensions— isolation and control of variables, combinatorial reasoning, correlational reasoning, probabilistic reasoning, and proportional reasoning (Han, 2013; Lawson, 1978). Based on these dimensions, Tobin



and Capie developed the Test of Logical Thinking (TOLT) in 1981, which was a tool to measure five dimensions of scientific reasoning presented by Lawson (1978). Roadrangka et al. (1983) developed the Group Assessment of Logical Thinking (GALT) using some aspects of Lawson's CTSR (1978) and Tobin and Capie's TOLT (1981). Utilizing the GALT method, researchers were able to measure six types of reasoning abilities—conservation, correlation, professional reflection, conductivity variables, probability reasoning, and combination reasoning. However, these early tools did not show meaningful correlation or prediction in solving problems requiring hypothesis-deductive reasoning (Niaz & Robinson, 1992).

In 2000, Lawson developed LCTSR, an upgraded inspection tool for measuring hypothesis-deductive thinking by modifying and supplementing the CTSR tool developed in 1978. In LCTSR, the combination reasoning in CTSR was replaced by correlation and hypothesized reasoning, and all questions were presented in the multiple-choice format. This LCTSR tool has been widely used to measure scientific reasoning ability (Han, 2013). However, researchers have suggested that LCTSR may have some problems. The tool is extremely isomorphic and highly dependent on questions. Besides, each of the 24 questions in the LCTSR tool are arranged in pairs; thus, there are 12 items, and only ten among them can be evaluated for formal reasoning ability (Kalinowski et al., 2019).

Kalinowski and Willoughby (2019) considered these problems as weaknesses of CTSR tools developed before modern psychometric methods were commonly used. Therefore, they reconstructed the test in the MSU Formal Reasoning Test (FORT) to increase the reliability of the test scores. The FORT tool comprises 20 questions, determined through a repetitive process of developing, conducting, analyzing, and modifying for reducing dependence on questions and increasing reliability. This tool was developed to measure five types of reasoning ability (COV, hypothesis testing, correlative reflection, probability, and professional reflection) while evaluating reasoning ability in a relatively short time (utilizing 99% of the 20 minutes students are provided) through a simple method (1 dimension instrument) (Kalinowski et al., 2019).

### *Factors Influencing Students' Scientific (Formal) Reasoning*

#### *Learning experience*

The importance of scientific reasoning ability has been consistently and extensively emphasized in science education standards and curricula as it has been regarded as a skill necessary for successfully conducting scientific exploration tasks (Han, 2013). In Piaget's model of cognitive development, young adults should have reached the final stage of maturation, thus, college students were expected to function at the formal operational level of cognitive development (Vass, Schiller, & Nappi, 2000). Theoretically, different learning experiences shaped by divergent cultures will be utilized to explain students' scientific reasoning performance and its association with epistemic beliefs in science (Yang et al., 2019). Few researchers have studied scientific reasoning ability in the Indonesian context. Yanto et al. (2009) suggested that lack of instruction emphasized on scientific reasoning skill and higher order thinking skills may influence low achievement in PISA.. Although various inquiry-based studies and instructions have been conducted in the Indonesian context (Effendi-Hasibuan & Mukminin, 2019; Hardianti & Kuswanto, 2017), their effect in improving scientific reasoning ability should be cautiously interpreted. The scientific reasoning ability formed through inquiry learning is considered much more important in Indonesian higher levels of education (Alameddinea & Ahwal, 2016). Therefore, given the upper level of cognitive development and the richer learning experience that university students have, they are expected to have higher scientific reasoning ability than the upper-secondary education students.

#### *Gender Difference*

Research studies focusing on gender differences in science education have been conducted since the 1960s to highlight the issue of inequality (Bianchini et al., 2000; Brotman & Moore, 2008). Furthermore, UNESCO (2017) reported the gender gap ratio in science research and noted that 70% science researchers were male. A report by PISA also highlighted gender differences to improve average performance and understand how students learn. Owing to the increased importance of gender equity in science, studies have noted the gender gap in the science education field, including academic self-efficacy (Huang, 2013; Sya'bandari et al., 2019), academic engagement and students' gender identity (Kessels et al., 2014), science self-concept (Jansen et al., 2014), and attitude toward science (Aini et al., 2019; Rusmana et al., 2021). It will also serve to understand academic achievement through scientific



reasoning ability and address quality and equity concerns. Some studies had noted that scientific reasoning was indicated to have no significant difference between males and females (Al-Zoubi et al., 2009; Dimitrov, 1999; Lappan, 2000; Valamides, 1996), meanwhile, others also reported a significant gender difference (Soyibo, 1999; Valanides, 1997; Yang, 2004). Severiens and ten Dam (1997) suggested the importance of investigating gender differences in order to be able to draw conclusions on the processes involved in learning. Thus, Yang (2004) compared differences in responses by males and females to examine possible gender effects on reasoning.

### *Metacognitive and confidence*

Reasoning is one of the high order thinking skills in the 21st century that are communicated by number sense and metacognition (Amin et al., 2020; Çekirdekci, Şengül, & Doğan, 2018). Recently there is a growing interest in metacognitive processes that accompany performance of scientific reasoning (Ackerman & Thompson, 2017) or critical thinking (Kelly & Ho, 2010). Ackerman and Thompson (2017) had noted that providing insight into control and monitoring metacognitive processes and their correlation with scientific reasoning was the goal of meta-reasoning study. Furthermore, reasoning was reported to have a positive correlation with self-efficacy of the students (Lawson et al., 2007). Kleitman and Stankov (2007) administered a confidence rating in each item in a test to indicate how confident the participant was that the chosen answer was correct. The similar procedure was also performed in other studies (Hwang et al., 2021; Lawson et al., 2007; Rusmana, Roshayanti, & Ha, 2020). Further, cognitive tests were also recognized along with confidence and metacognitive measurements (Kleitman & Stankov, 2007). In the examination, they noticed that metacognitive inventories afforded unique variance in confidence scores when cognitive abilities were controlled (Fritzsche et al., 2012).

### *Research Aims and Research Questions*

According to the background and the explanation above, this study primarily examines the scientific reasoning of Indonesian upper-secondary school and university students. It also compares the scientific reasoning ability of students in Indonesia and the US and investigates gender differences and academic level interactions impacting their ability. Furthermore, the results of a distractor analysis for scientific reasoning in Indonesian students were obtained. The research questions explored in this study were as follows:

RQ1: What is the status of scientific reasoning ability in Indonesian students, compared with US students?

RQ2: What are the alternative conceptions of Indonesian students in response to FORT items?

## **Research Methodology**

### *General Background*

This study used quantitative methods by using scientific reasoning instruments and questionnaires to obtain scientific reasoning ability and confidence level data from university and upper-secondary school students in West Java Province, Indonesia. The survey was applied in the academic year 2020.

### *Participants*

Data were collected from 372 university students and 528 upper-secondary school students in Indonesia through a convenience sampling method (Cohen et al., 2007). Following the ethical guidelines for educational research, at the beginning of data collection, the participants' voluntary informed consent to be involved in the study was obtained. They were given information about research, who conducted the research, why their participation is necessary, what they will be asked to do, and how the information they provided will be used. They were also informed that they may request feedback about their reasoning score if they need it. The participants were gathered from three private and public upper-secondary schools, and one university in west Java Province. Of a total of 900 participants, 276 (30.7%) participants were male and 618 (68.7%) were female. Six students did not provide complete information regarding gender (missing system). The university students were enrolled in science-related majors, such as biology, biology education, physics, physics education, and science education.



*Instrument and Procedures*

The scientific (formal) reasoning test (hereafter, FORT instrument) (Kalinowski & Willoughby, 2019) comprises 20 items for measuring students' reasoning ability and was used as the primary research instrument. Furthermore, a 5-point Likert scale (1 = "strongly not confident" and 5 = "strongly confident") was used to measure students' confidence in the correctness of their answer. The instruments were translated into Indonesian by experts in both English and Indonesian to ensure the context of the questions remained unchanged. Furthermore, students filled a 19 items questionnaire of the Metacognitive Awareness Inventory (MAI). It measured students' metacognition in the framework of knowledge and cognitive regulation. The shortened version of MAI reported to have better model function and validity of scoring inference (Harrison & Vallin, 2017). Students also filled their self-reported rank in the answer sheet provided. The three-parameter (3PL) item response theory was used to provide validity evidence (Table 1), item difficulty, and person ability. Cronbach's alpha was calculated to determine internal consistency (Cronbach's alpha = .65).

**Table 1**  
*Psychometrics Properties of FORT Instrument*

Construct	Item Number	Item Difficulty	Discrimination	Guessing Rate
Control of Variable	1	2.601	1.789	0.000
	7	0.297	1.373	0.002
	16	1.173	0.460	0.000
	20	0.509	1.319	0.002
Hypothesis Testing	2	1.320	0.767	0.000
	5	0.468	0.719	0.001
	9	1.747	-0.268	0.000
	12	1.374	0.287	0.000
Correlation	4	1.236	0.159	0.000
	10	0.835	-0.430	0.001
	13	1.627	0.158	0.000
	18	1.227	0.588	0.001
Proportions	6	0.572	1.360	0.004
	11	1.011	1.923	0.000
	15	1.376	-0.254	0.001
	17	-0.721	1.558	0.004
Probability	3	0.565	1.796	0.014
	8	1.202	1.177	0.008
	14	1.547	0.556	0.003
	19	0.783	1.224	0.005

*Note:* Five types of formal reasoning of the 20 questions on the FORT Instrument: COV, hypothesis testing, correlational reasoning, probabilistic thinking, and proportional thinking. Item validation and measurement are calculated based on 3PL IRT: the difficulty, discrimination, and guessing rate of the question. Negative difficulty indicates that the question is easier than the average, and positive difficulty indicates that the questions are more difficult. High discrimination values indicate how well the question differentiates between high- and low-ability students. A zero-guessing rate occurs when a low-ability student always selects the wrong answer (Kalinowski, 2019). Item\_17 and item\_1 are considered as the least and the most difficult items. Moreover, three items (item\_9, item\_10, and item\_15) are calculated lower than zero in discrimination.



Data Analysis

To compare the scientific reasoning ability of students in Indonesia and the US, the proportion of students who answered questions correctly were calculated ( $P_{Correct}$ ). This proportion was also calculated among students in each academic level and gender. Furthermore, students' person ability generated 3PL used for further analysis. This person's abilities measured students' performance on scientific reasoning (reasoning ability), with zero values as an average score from overall populations. Two-way ANOVA was performed to compare mean differences between groups, based on academic level and gender, and to observe interactions between the variables. To estimate the relation between each variable to students' reasoning scores, regression analysis was performed. In this study, confidence level, metacognitive score, and students' self-ranking were used as covariates. Furthermore, the distractor options within each multiple-choice question were investigated. Students' mean score of confidence level in selecting the correct answer and distractor was analyzed using an independent t-test. Furthermore, characteristics of student scientific reasoning were grouped and analyzed through clustering analysis.

Research Results

Scientific Reasoning Ability of Indonesian Students with Regard to US Data, Gender, and Academic Level

The scientific reasoning ability of Indonesian students in the five constructs was separated based on academic level and gender. Following the results in Table 2 showed the proportion of students answering the item correctly within Indonesian students compared to the US students ( $P_{Correct}$ ). Overall, the college students in the US considerably have a higher proportion of correct answers in all constructs, compared with students in Indonesia. Among Indonesian participants, university students primarily have a higher proportion of correct answers, compared with upper-secondary school students. In terms of gender, male participants noted a higher proportion of correct answers in all constructs for both university and upper-secondary school, excluding the scores of upper-secondary school students when responding to the probability construct wherein both genders have a similar proportion of correct answers (0.26).

In addition, the comparison of reasoning ability within academic levels and gender was analyzed by two-way ANOVA. The results show that academic level significantly impacts the reasoning ability ( $F [1, 887] = 126.946, p = <.001, PES = 0.125$ ) which university students have significantly higher reasoning ability than high school students (Seen in Figure 1). Gender also contributes to the significant difference of reasoning ability ( $F [1, 887] = 11.489, p = 0.001, PES = 0.013$ ), wherein male students show significantly higher reasoning ability, compared with their female counterparts. Finally, the interaction of academic level and gender denotes significant differences in the reasoning ability ( $F [1, 887] = 6.982, p = 0.008, PES = 0.008$ ). Figure 1 shows that at the high school level, male students have slightly higher reasoning ability, compared with female students while at the university level, significant differences are found in the reasoning ability of male and female students.

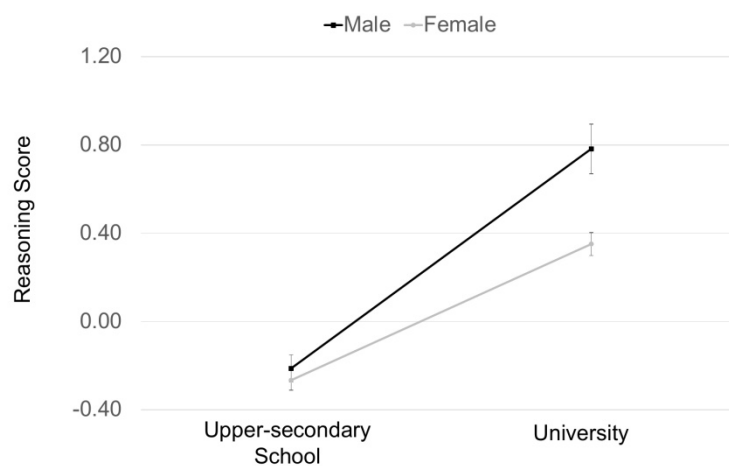
**Table 2**  
Comparison of Scientific Reasoning Ability between Students in Indonesia and the US

Construct	Item Number	PCorrect USA	$P_{Correct}$ Indonesia			
			University		Upper-secondary School	
			Male	Female	Male	Female
	1	0.71	0.37	0.21	0.06	0.04
	7	0.73	0.58	0.62	0.35	0.28
Control of Variable	16	0.49	0.32	0.26	0.24	0.21
	20	0.74	0.66	0.51	0.24	0.32
	Mean	0.67	0.48	0.40	0.22	0.21



Construct	Item Number	PCorrect USA	$P_{Correct}$ Indonesia			
			University		Upper-secondary School	
			Male	Female	Male	Female
Hypothesis Testing	2	0.58	0.39	0.30	0.17	0.16
	5	0.78	0.50	0.52	0.30	0.30
	9	0.21	0.15	0.13	0.21	0.14
	12	0.47	0.23	0.17	0.23	0.22
	Mean	0.51	0.32	0.28	0.23	0.20
Correlation	4	0.44	0.31	0.24	0.26	0.17
	10	0.43	0.24	0.26	0.40	0.31
	13	0.23	0.16	0.14	0.21	0.16
	18	0.48	0.27	0.29	0.23	0.17
	Mean	0.40	0.25	0.23	0.28	0.20
Proportions	6	0.66	0.71	0.41	0.34	0.32
	11	0.48	0.48	0.43	0.29	0.21
	15	0.15	0.19	0.17	0.24	0.21
	17	0.82	0.82	0.73	0.56	0.54
	Mean	0.53	0.55	0.44	0.36	0.32
Probability	3	0.81	0.60	0.45	0.34	0.33
	8	0.68	0.36	0.29	0.25	0.22
	14	0.48	0.34	0.19	0.15	0.17
	19	0.61	0.52	0.35	0.29	0.32
	Mean	0.65	0.45	0.32	0.26	0.26

**Figure 1**  
The Relationship of Academic Level and Gender with Students' Reasoning Ability



In this study, confidence level, metacognitive score, and students' self-ranking are considered covariates. These variables are expected to influence the effect of gender and academic level on reasoning ability. Thus, regression analysis was performed. Before the adjustment, academic level and gender were associated significantly with

reasoning ability ( $\beta = 0.358, p = <.001, \beta = 0.118, p = .002$ , respectively). The positive value of  $\beta$  coefficient shows that students with higher academic levels derive higher reasoning ability. It also denotes that male students have higher ability than female students. These results corroborated the result of two-way ANOVA. After adjusting the covariates, slightly different values of  $\beta$  coefficient are noted for the association of academic level and gender ( $\beta = 0.362, p = <.001, \beta = 0.102, p = .006$ , respectively). This means that adding covariates did not significantly change the results, instead having the more accurate data interpretation.

Furthermore, particular items that establish the highest gap of scientific reasoning ability between the groups were identified. The highest gap score between university students in the US and Indonesia is found in the first item of the "COV" construct, followed by the eighth item of "probability construct." The seventh and twentieth items of "COV" create the gap between university and upper-secondary school students' reasoning ability in Indonesia. Additionally, the highest gap between Indonesian male and female students' scientific reasoning ability was noted in the fourth and tenth items of correlation construct. Of the six items that established the biggest gap ability, students' answer to the questions was found to be patterned. In some items, instead of selecting the correct option, several students were distracted to select a particular option. For instance, in the first item (shown in appendix) that inquired about the best soil type to grow corn, instead of choosing the D option as the correct answer, most students were distracted by the C option. The circumstance is unique to be explored, thus students' alternative questions by analyzing distractor options in particular items were identified.

In addition, Indonesian student classification based on scientific reasoning outcome in each construct was performed by using mclust package from R. Referring to Fraley & Raftery (2002) the best number of clusters can be designated based on the highest BIC from clustering possibility. Analysis of the data revealed that clustering possibility of 1—20 groups results in the highest BIC on clusters of 13 groups (BIC=-10977.21). However, the cluster of 13 groups was found to be insubstantial because of a highly unequal distribution. The analysis with three clusters (BIC=-12541.32, VEE model: ellipsoidal, equal shape and orientation and variable volume) was run and was able to give meaningful results and unique characteristics in each scientific reasoning construct as illustrated in Figure 2.

**Figure 2**  
*Clustering Analysis of Scientific Reasoning Outcome*

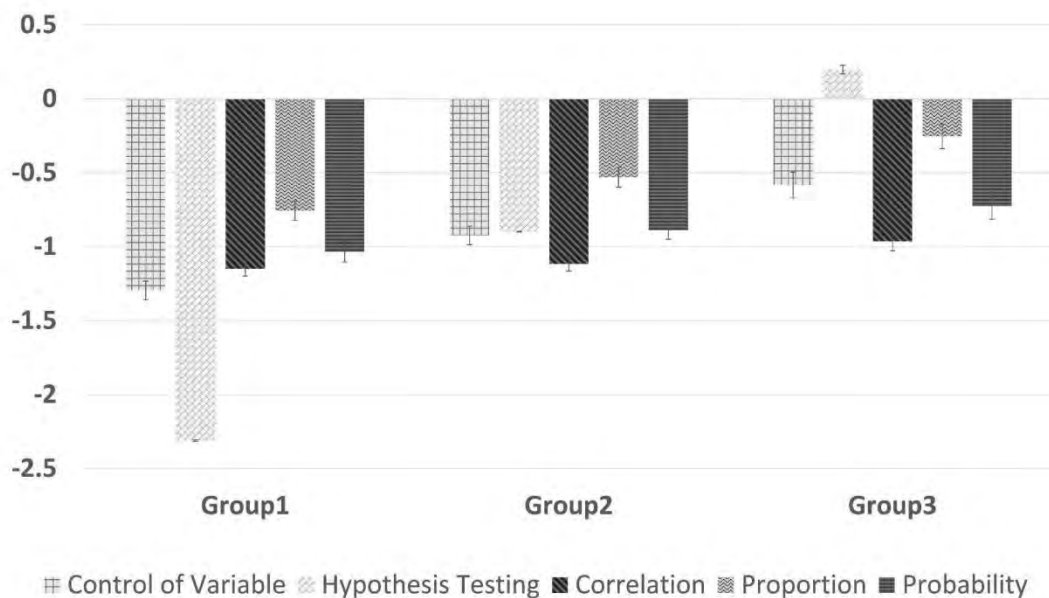


Figure 2 represents the groups with the least to the highest mean value on each scientific reasoning construct. Group 1 consists of 373 participants (41.5%) who had the lowest mean value in each scientific reasoning construct. In this group, the lowest mean value is on the hypothesis testing dimension ( $M=-2.31, SD=0.094$ ) and the highest mean value is on the proportion construct ( $M=-0.76, SD=1.130$ ). Group 2 consists of 301 participants (33.4%) with





mean value in proportion construct ( $M=-0.53, SD=1.297$ ) was found to be the leading construct. Group 3 was composed by 226 participants (25.1%). It has the highest mean value of reasoning among all groups, especially in the hypothesis testing construct ( $M=0.19, SD=0.428$ ). According to Figure 2, it is noticeable that there is a significant difference in hypothesis testing construct between the three clusters. Table 3 below represents the distribution of each group based on academic level and gender.

**Table 3**  
*Students' Distribution on Clustering Analysis*

Group	Gender				Academic			
	Male	%	Female	%	University	%	Upper-secondary School	%
G1	94	34.1	276	44.7	165	44.4	208	39.4
G2	104	37.7	194	31.4	91	24.5	210	39.8
G3	78	28.3	148	23.9	116	31.2	110	20.8
Total	276	100	618	100	372	100	528	100

In furtherance of identifying the group characteristics based on demographic variables, chi-squared tests were performed. The significant relation was found between the groups and gender ( $\chi^2(8.850), p = .012$ ). The similar founding was also encountered between groups and the academic level characteristics ( $\chi^2(25.901), p < .001$ ). Group 1 which is found to be the group with the least mean value of scientific reasoning in all constructs has the highest number of participants and it is dominated by female students (44.7%), meanwhile, Group 3 that has the leading mean value in all scientific reasoning constructs has the least distributed students which are dominated mostly by male students (28.3%).

#### *Alternative Conceptions of Indonesian Students in Response to FORT Items*

Following up on the results of RQ1, many Indonesian students selected certain distractors with relatively high confidence. In particular, although the option was wrong, most students believed it was right. Thus, this study explored students' alternative conception in response to the FORT items through the analyses of distractors and confidence levels. Table 4 presents the summary of Indonesian students' alternative conception in response to the FORT items. It presents the correct answer of FORT (correct answer), Most Chosen Distractor (MCD), the proportion of students answering the item correctly ( $P_{\text{Correct}}$ ), proportion of students selecting the distractor ( $P_{\text{Distractor}}$ ), and the mean comparison of confidence level for selecting between the correct answer and distractor as the correct answer. Finally, the results are separated based on the FORT constructs.

Among 20 items, the proportion value of the correct answer is greater than the correct distractor in only four items (item\_7, item\_5, item\_17, and item\_3). In general, students selected the distractor option with relatively low confidence. Nonetheless, as shown in Table 4, from 8 items showing a significant difference in confidence level, there are two items wherein the confidence level in selecting the distractor significantly outweighed the correct option. They are item\_1 in COV construct ( $t = -2.419, p = .016, d = 0.250$ ) and item\_13 in correlation construct ( $t = -3.986, p < .001, d = 0.373$ ). These items are presented in the Appendix and will be further discussed.

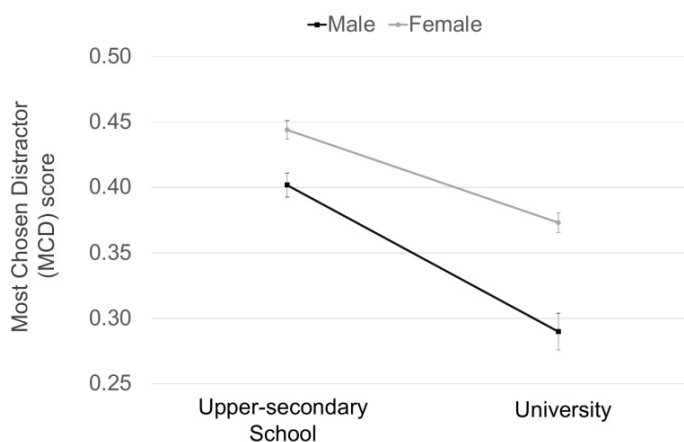
Furthermore, the most chosen distractor (MCD) score was calculated. The relations of MCD scores were analyzed by considering the effect of academic level and gender identity. The data shows a statistically significant difference of MCD score on academic levels ( $F[1, 890] = 73.584, p < .001, PES = 0.076$ ). Particularly, Figure 3 indicates that university students have significantly lower MCD scores than upper-secondary school students. In addition, the MCD score is also significantly different, based on gender ( $F[1, 890] = 34.659, p < .001, PES = 0.037$ ). Male students have a significantly lower MCD score than female students. The interaction of academic level and gender is not significantly different, compared with the MCD score ( $F[1, 890] = 3.713, p = .054, PES = 0.004$ ).



**Table 4**  
*Indonesian Students' Alternative Conceptions in the FORT Items*

Construct	Item Number	Correct answer	MCD	P <sub>Correct</sub>	P <sub>Distractor</sub>	Confidence level		
						t	p	d
Control of Variable	1	D	C	0.12	0.80	-2.419	.016	0.250
	7	C	B	0.43	0.30	6.338	<.001	0.506
	16	A	B	0.24	0.26	0.657	.512	0.063
	20	A	D	0.39	0.42	0.445	.657	0.033
Hypothesis Testing	2	B	C	0.22	0.35	0.637	.524	0.058
	5	D	B	0.39	0.22	-0.077	.939	0.007
	9	B	C	0.15	0.39	-1.148	.252	0.117
	12	B	E	0.20	0.43	0.461	.645	0.042
Correlation	4	C	B	0.23	0.52	-0.152	.880	0.013
	10	A	B	0.31	0.35	0.500	.617	0.041
	13	B	A	0.17	0.62	-3.986	<.001	0.373
	18	D	B	0.23	0.42	0.912	.362	0.079
Proportions	6	B	C	0.38	0.51	0.347	.729	0.025
	11	C	D	0.32	0.53	3.064	.002	0.230
	15	A	C	0.20	0.43	-0.203	.839	0.018
	17	A	B	0.63	0.15	3.834	<.001	0.369
Probability	3	A	C	0.39	0.29	3.803	<.001	0.312
	8	C	E	0.26	0.28	2.524	.012	0.232
	14	B	C	0.18	0.26	0.525	.600	0.054
	19	A	C	0.34	0.43	6.660	.0001	0.517

**Figure 3**  
*The Relationship of Academic Level and Gender with MCD Score*



The association between academic level and gender on the MCD score while considering the confidence level, metacognitive score, and self-ranking were analyzed. Before the adjustment, academic level and gender are



significantly associated with MCD score ( $\beta = -0.278, p = <.001, \beta = -0.164, p = <.001$ , respectively). After accounting for the covariates, a difference is noted in the value of coefficient  $\beta$  for both academic level and gender ( $\beta = -0.272, p = <.001, \beta = -0.184, p = <.001$ , respectively). A slightly different score demonstrated the more accurate data without significantly revealing the different result after the adjustment of covariates. The negative value shows that students with increased academic level reported having lower distractors. Moreover, male students are expected to have lower distractions than female students. These results also corroborated the two-way ANOVA results.

## Discussion

### *The Status of Scientific Reasoning Ability of Indonesian Students*

In this study, the proportion of correct answers from Indonesian students in five constructs of scientific reasoning (COV, hypothesis testing, correlation, proportions, and probability) were examined, then, compared to that of students in the US (Kalinowski & Willoughby, 2019). The results show that students in the US had higher correct answers in all constructs, compared with Indonesian students. The result is in line with the low performance of Indonesian students in the international standardized tests that examine students' science reasoning, such as Trends in International Mathematics and Science Study (TIMSS). According to Martin et al. (2016), the science cognitive domain being tested in TIMSS covers the knowing, applying, and reasoning domain. The most recent result of TIMSS in 2015 showed that Indonesia ranked 54 out of 57; Indonesia was ranked fourth from the bottom. Indonesian students' performance in all science cognitive domains (average scale score is 397) is lower than the international average benchmark (500). In contrast, students' performance in the US was placed in the top ten of high achievers. While Indonesian students' reasoning scores were lowest in the knowing, applying, and reasoning domains which are at 397, 392, and 390, respectively, the American students' reasoning score in TIMSS was above average, counted for 542 (Martin et al., 2016). The findings were interpreted by analyzing the emphasis of scientific reasoning in the science curriculum of two countries. In terms of the framework for K-12 science education, students in the US explicitly learn scientific reasoning in science learning because it is a key element of science (Gross, 2011). Additionally, in the new K-12 science education standards, reasoning and offering solutions is one of the eight essential elements included in the curriculum (NRC, 2012). While the US science curriculum has emphasized the fusion of scientific reasoning as a part of science instruction; Novia and Riandi (2017) suggested that science learning in Indonesia has not appropriately facilitated the development of scientific reasoning. Although reasoning is one of five scientific approaches (observing, questioning, experimenting, reasoning, and communicating the result) promoted in the new curriculum of 2013 in Indonesia, Suyanto (2018) pointed out that reasoning was the least appropriate scientific approach implemented in the classroom. Some researchers (Rustaman, 2009; Utomo et al., 2018; Wasis, 2014) noted that Indonesian students rarely practice their reasoning in science instruction. Students more often learn science in conventional ways, such as lecturing, so that they are barely challenged to express their ideas and indulge in problem-solving and decision-making. Moreover, Indonesian students are rarely given higher order thinking skills (HOTS) test items; the type of questions in the assessment mostly only required students to memorize content (Novia & Riandi, 2017). Among 20 FORT items answered by the students, two resulted in a notable gap between the reasoning scores of students in Indonesia and the US: one item is in the "COV" construct (item no. 1) and the second item is in the "probability" construct (item no. 8). Presented in the Appendix, these two items at first glance are considered simple to answer, thereby leading students to choose the wrong option. Indonesian students thus face difficulties answering such items that require higher order thinking skills and involve the process of in-depth analysis and evaluation (Utomo et al., 2018).

Regarding the comparison of scientific reasoning between university and upper-secondary school students, our study showed that in most constructs, the reasoning ability of university students was significantly higher, compared with that of upper-secondary school students. This finding supports the common perspective that as students pursue higher education, they are expected to perform a higher level of scientific reasoning (Ding et al., 2016). Some researchers (Fencl, 2010; Klahr et al., 2011; Kuhn, 2002; Nieminen et al., 2012) have agreed that scientific reasoning skills are transferable and can be trained through education; therefore, it is reasonable that students' scientific reasoning progresses and proceeds in the higher level of education as they experience learning progression. Ding (2018) noted that the design of science instruction may influence the progressive development of scientific reasoning among university students. According to Fencl (2010) and Gerber et al. (2001), the process of teaching and learning that engage students in scientific inquiries (e.g., designing experiments, generating hypoth-



eses, making inferences, and participating in evidence-based arguments) can improve student scientific reasoning skills that such activities are massively promoted among undergraduate students, consistent with the high value given to these skills in the tertiary level (Star & Hammer, 2008). Additionally, increased experiences of formal and informal learning for university students can contribute to the advancement of their reasoning (Gerber et al., 2001).

Furthermore, the items that contribute to the notable gap between upper-secondary school and undergraduate students are primarily in the "COV" dimension (the seventh and twentieth items), thereby indicating that such items were perceived to be more difficult for upper-secondary school students. As a core sub-skill of scientific reasoning, COV refers to the ability to scientifically manipulate experimental settings (some variables are kept constant while others are changed) during data collection (Zhou et al., 2016). The developmental stages of COV skills are starting from controlling a few variables to complex multivariable and causal analysis. Chen and Klahr (1999) utilized these skills to develop higher academic levels. Consider the twentieth COV item (depicted in Appendix) for an example wherein the item requires students to deal with some variables (presence of light, number of fish, and temperature) and analyze the cause-and-effect relationship of those variables. Given that the item probably requires intermediate to high-end skills and the nature of FORT items that are designed for university students, it may explain the difficulties encountered by upper-secondary school students while answering such questions.

In terms of gender difference in scientific reasoning, the results showed that the scores of male students are higher than that of female students, as in the findings of previous studies (Coletta et al., 2012; Liben & Golbeck, 1980; Valanides, 1997; Yang, 2004). The distribution from the clustering analysis also shows that female students dominate the least achiever group. Baron-Cohen (2003) explained that male students show good performance when learning objects and mechanical relationships, while female students show better performance when learning about people, emotions, and personal relationships. Additionally, Ward and King (2020) mentioned that female students tend to be more intuitive in their responses toward uncertainty. Consequently, true scientific reasoning that requires an understanding of the physical world, rather than common sense knowledge about objects, provides advantages for male students (Spelke, 2005). The FORT items that favor male students are included in the correlation construct and are consistent with the findings of Mari (2012). Correlational reasoning refers to the ability to determine the strength of mutual or reciprocal relationships among variables (Lawson et al., 1979; Piraksa et al., 2014). Correlational reasoning involves the ability to make inferences; thus, male students with a more rational and analytical perspective experience advantages in determining associations among two or more variables.

According to clustering analysis in each scientific reasoning dimension that has been shown in Figure 2, the Hypothesis Testing dimension denotes a contrast gap in each group. Some authors (Inhelder et al., 1958; Neimark, et al., 1975) argue that the ability to test hypotheses has been postulated to play a central role in a variety of cognitive processes. As noted by Khun (1989), the heart of scientific thinking is the coordination of theories and evidence. Furthermore, the most central, essential, and general skills that define scientific thinking considerably consist of: (1) the skills to consciously articulate accepted theory, (2) to know which evidences support or contradict the theory, and (3) to justify why the coordination of available theories and evidence could lead to the acceptance of the theory (Kuhn, 1989).

#### *Students' Alternative Conceptions in Response to FORT Items Based on Distractor Analysis*

Among 20 FORT items, 16 weigh higher for the distractor option, instead of the correct option. This result shows that a higher proportion of Indonesian students selected distractors as their answer. Particularly, there were two items related to the skill of causal reasoning wherein students believed that the distractor was the correct answer. This fact was demonstrated by the significant difference in confidence level between the distractor and correct option. Firstly, item 1 in the COV construct. Woolley et al. (2018) reported fallacies in scientific reasoning that commonly occur among undergraduate students—identifying and controlling variables. Additionally, middle school students have falsely identified and controlled variables because they could not apply those variables in different contexts, although students understand the meaning of the independent and dependent variables (Leatham, 2012). Kuhn (2007) stated that students often experienced difficulties in the reasoning of multiple variables. Item 1 presented two figures and asked students about the significant variables affecting the experiment (see Appendix). Although there was no control variable, most students compared and selected soil B and fertilizer Y as significant variables in the experiment. Most students believed that option C "Soil B is best for growing her corn, and Fertilizer Y is best for growing her corn" is the correct answer. Based on this case, it can be assumed that students only observe the number of crops in the figures and proceed with an automatic and quick mental process of perception and



memory with little effort of thinking. In this case, it is assumed students used a system of thinking 1 (intuitive) and generated a fast sense of incoming information by integrating it based on their background knowledge and beliefs (Kahneman, 2011). When students believe their initial conclusion to be true, they will also believe that the arguments support it, even if it is false (i.e., confirmation bias and false certainty error; Covitt et al., 2013). This process also leads students to have an overconfident bias where they erroneously believe the answer is correct. This type of thinking discourages people from reasoning in everyday life. Meanwhile, interpreting the most significant variables in the experiment requires deliberate and effortful reasoning.

Second, item 13 from the construct of correlational reasoning. According to Pohl and Pohl (2004), the assessment of the correlations is a crucial subject of adaptive intelligence and behavior. Humans often assess correlation inaccurately and irrelevantly, due to the cognitive shortcut to make a fast judgment or *heuristics*. Specifically, this element is called an *illusory correlation*, and it occurs because students overemphasize one outcome that they perceived to be important and is easy to recall, thereby leading them to underestimate the other outcomes. To determine hidden assumptions leading to the bias correlation, the contingency table test is used, and this is presented in item 13 (see Appendix). The result revealed that students reflect on the unequal weighting of the four cells. Owing to the asymmetry to the present and absent factors (feature positive effect), higher frequency in the cell determines the weigh-in of the correlation assessment. Therefore, more weight was given to present HDLs (cause) and high blood pressure (effect), instead of the absent cases and omitted ones. This step is followed by correlation's cell of present HDLs and no high blood pressure, and correlation's cell of absent HDLs and high blood pressure. Moreover, the least weight was given to the cell of absent HDLs and no high blood pressure. Observing the present HDLs (cause) and high blood pressure (effect) with 150 participants encourage students to weigh higher than other cells with frequency 50, 30, and 10. Particularly, the question asks directly about the presence of the cause and effect of the correlation "Does there appear to be an association between HDLs and high blood pressure?" Therefore, most students believed that there was an association between HDLs and high blood pressure. The emergence of the illusory correlation is the result of the incomplete learning to question the relation of the groups with frequent and infrequent traits (Murphy et al., 2011). In social psychology, the perceived correlation of characteristics and group membership allows for stereotypical thinking (Berndsen et al., 2001; Pohl & Pohl, 2004).

The two items illustrate how students experienced false reasoning due to the errors in cognition. Kahneman (2011) stated that cognitive bias cannot be avoided in our daily life because humans prefer to think simply. Mindfulness and deliberate thoughts are necessary while making decisions (reflective thinking). Furthermore, the result of this study showed that students with higher academic levels are lesser in selecting distractors. This finding is in accordance with that of Ding et al. (2016) that students with higher educational levels are estimated to have a higher level of reasoning, despite a minor variation within the college group across the entire 4 years. With regard to the gender difference, female students choose more distractors than males. Female students had more intuitive responses to the topic that is uncertain and has not to accept the empirical attention comprehensively (Frederick, 2005; Pennycook et al., 2016; Toplak et al., 2011; Ward & King, 2020). Females are reportedly more emotional than males and rely on intuition (Ward & King, 2018). On the contrary, males show more rational and analytical thinking in decision-making (Campitelli & Gerrans, 2014; Rogers et al., 2019; Sladek et al., 2010).

The FORT was able to measure students' scientific reasoning ability. Through distractor analysis, this study has introduced a further approach that can enrich and deepen analysis in the use of FORT instruments. In order to change the alternative conceptions, Lawson (2003) indicated the relationship between students' alternative conceptions and reasoning ability as the study pointed out that students needed to be aware of their alternative conceptions and scientific conceptions, together with the evidence and reasoning that validate the alternative conceptions. Furthermore, information about Indonesian students' responses to scientific reasoning in the perspective of gender and academic level can be a germane literature for the future work in scientific reasoning using FORT instrument.

## Conclusions and Implications

This study considers scientific reasoning ability within Indonesian upper-secondary school and university students. In addition, a comparison with the scientific reasoning ability of students in the US is also investigated, as well as an analysis of gender, confidence level, and metacognition. By testing 900 participants of Indonesian upper-secondary school and university students using the FORT, Indonesian students were found to have lower scientific reasoning ability than American. This result also denotes a significant difference of scientific reasoning ability between Indonesian and American students where a correspondence result is also observed in PISA and



TIMSS assessment. With regards to the relation of academic level and gender, it is found that male upper-secondary school students had slightly higher reasoning ability while, male undergraduate students achieved a significantly higher reasoning ability than their female counterparts. The higher scientific reasoning ability in higher academic level is also noted. The university students selects lesser distractors than upper-secondary school students. Owing to the Indonesian students' low achievement in scientific reasoning, science instruction in Indonesia rarely accommodates scientific reasoning practice. In terms of Indonesian students' alternative conception, it is noted that most students choose more distractors and among them, female students were recorded to choose more distractors with a high confidence level. Thus, our findings show that a false initial conclusion and false supporting arguments may lead to an overconfident bias, where students believe that the answer is correct although it is wrong. In addition, students perform automatic and quick mental processes of perception and memory with little effort of thinking that is connected to intuitive thinking. Meanwhile, slow, deliberate, and effortful reasoning is necessary to interpret the most significant variables in this experiment. Hence, becoming familiar with scientific reasoning through a constant practice of high order thinking skills or reasoning assessment during learning is suggested for students.

### Declaration of Interest

Authors declare no competing interest.

### References

- Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, 21(8), 607-617. <https://doi.org/10.1016/j.tics.2017.05.004>
- Aini, R. Q., Rachmatullah, A., & Ha, M. (2019). Indonesian Primary school and middle school students' attitudes toward science: Focus on gender and academic level. *Journal of Baltic Science Education*, 18(5), 654-667. <https://dx.doi.org/10.33225/jbse/19.18.654>
- Al-Zoubi, T., El-Shara, I., & Al-Salam, M. K. (2009). The scientific reasoning level of students in the faculty of science in al-husseini bin talal university and its affection of gender, teaching level, and specialization. *An-Najah University Journal for Research-Humanities*, 23(2), 401-437. <http://hdl.handle.net/20.500.11888/2348>
- Alameddine, M. M., & Ahwal, H. W. (2016). Inquiry based teaching in literature classrooms. *Procedia-Social and Behavioral Sciences*, 232, 332-337. <https://doi.org/10.1016/j.sbspro.2016.10.031>
- Allchin, D., & Zemplén, G. Á. (2020). Finding the place of argumentation in science education: Epistemics and Whole Science. *Science Education*, 104(5), 907-933. <https://doi.org/10.1119/1.2976334>
- Amin, A. M., Corebima, A. D., Zubaidah, S., & Mahanal, S. (2020). The correlation between metacognitive skills and critical thinking skills at the implementation of four different learning strategies in animal physiology lectures. *European Journal of Educational Research*, 9(1), 143-163. <https://doi.org/10.12973/eu-jer.9.1.143>
- Bao, L., Fang, K., Cai, T., Wang, J., Yang, L., Cui, L., ... & Luo, Y. (2009). Learning of content knowledge and development of scientific reasoning ability: A cross culture comparison. *American Journal of Physics*, 77(12), 1118-1123. <https://doi.org/10.1119/1.2976334>
- Baron-Cohen, S. (2003). *The essential difference: The truth about the male and female brain*. Basic Books.
- Berndsen, M., McGarty, C., Van der Pligt, J., & Spears, R. (2001). Meaning-seeking in the illusory correlation paradigm: The active role of participants in the categorization process. *British Journal of Social Psychology*, 40(2), 209-233. <https://doi.org/10.1348/014466601164821>
- Bianchini, J. A., Cavazos, L. M., & Helms, J. V. (2000). From professional lives to inclusive practice: Science teachers and scientists' views of gender and ethnicity in science education. *Journal of Research in Science Teaching*, 37(6), 511-547. [https://doi.org/10.1002/1098-2736\(200008\)37:6<511::AID-TEA2>3.0.CO;2-3](https://doi.org/10.1002/1098-2736(200008)37:6<511::AID-TEA2>3.0.CO;2-3)
- Brotman, J. S., & Moore, F. M. (2008). Girls and science: A review of four themes in the science education literature. *Journal of Research in Science Teaching*, 45(9), 971-1002. <https://doi.org/10.1002/tea.20241>
- Burmester, M. A. (1952). Behavior involved in the critical aspects of scientific thinking. *Science Education*, 36(5), 259-263. <https://doi.org/10.1002/sce.3730360502>
- Burney, G. M. (1974). *The construction and validation of an objective formal-reasoning instrument*. University of Northern Colorado (Doctoral dissertation).
- Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, 42(3), 434-447. <https://doi.org/10.3758/s13421-013-0367-9>
- Čavojová, V., Šrol, J., & Jurkovič, M. (2020). Why should we try to think like scientists? Scientific reasoning and susceptibility to epistemically suspect beliefs and cognitive biases. *Applied Cognitive Psychology*, 34(1), 85-95. <https://doi.org/10.1002/acp.3595>
- Çekirdekci, S., Şengül, S., & Doğan, M.C. (2018). The relationship between number sense and metacognition. *Uluslararası Avrasya Sosyal Bilimler Dergisi*, 9(34), 2465-2481
- Chamberlin, T. C. (1898). The influence of great epochs of limestone formation upon the constitution of the atmosphere. *The Journal of Geology*, 6(6), 609-621
- Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, 70(5), 1098-1120. <https://doi.org/10.1111/1467-8624.00081>



- Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education*. Abingdon, OX: Routledge.
- Coletta, V. P., Phillips, J. A., & Steinert, J. (2012, February). FCI normalized gain, scientific reasoning ability, thinking in physics, and gender effects. In AIP conference proceedings (Vol. 1413, No. 1, pp. 23-26). American Institute of Physics. <https://doi.org/10.1063/1.3679984>
- Covitt, B. A., Harris, C. B., & Anderson, C. W. (2013). Evaluating scientific arguments with slow thinking. *Science Scope*, 37(3), 44-52.
- Dimitrov, D. M. (1999). Gender differences in science achievement: Differential effect of ability, response format, and strands of learning outcomes. *School Science and Mathematics*, 99(8), 445-450. <https://doi.org/10.1111/j.1949-8594.1999.tb17507.x>
- Ding, L., Wei, X., & Mollohan, K. (2016). Does higher education improve student scientific reasoning skills? *International Journal of Science and Mathematics Education*, 14(4), 619-634. <https://doi.org/10.1007/s10763-014-9597-y>
- Ding, L. (2018). Progression trend of scientific reasoning from elementary school to university: A large-scale cross-grade survey among Chinese students. *International Journal of Science and Mathematics Education*, 16(8), 1479-1498. <https://doi.org/10.1007/s10763-017-9844-0>
- Douglas, M., Wilson, J., & Ennis, S. (2012). Multiple-choice question tests: a convenient, flexible and effective learning tool? A case study. *Innovations in Education and Teaching International*, 49(2), 111-121. <https://doi.org/10.1080/14703297.2012.677596>
- Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data—evolution, challenges and research agenda. *International Journal of Information Management*, 48, 63-71. <https://doi.org/10.1016/j.ijinfomgt.2019.01.021>
- Effendi-Hasibuan, M. H., & MUKMININ, A. (2019). The inquiry-based teaching instruction (IbTI) in Indonesian secondary education: What makes science teachers successful enact the curriculum?. *Journal of Turkish Science Education*, 16(1), 18-33. <http://dx.doi.org/10.12973/tused.10263a>
- Fencl, H. S. (2010). Development of students' critical-reasoning skills through content-focused activities in a general education course. *Journal of College Science Teaching*, 39(5), 56-62.
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 611-631. <https://doi.org/10.1198/016214502760047131>
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4), 25-42. <https://doi.org/10.1257/089533005775196732>
- Fritzsche, E. S., Kröner, S., Dresel, M., Kopp, B., & Martschinke, S. (2012). Confidence scores as measures of metacognitive monitoring in primary students? (Limited) validity in predicting academic achievement and the mediating role of self-concept. *Journal for Educational Research Online*, 4(2), 120-142.
- Gerber, B. L., Cavallo, A. M. L., & Marek, E. A. (2001). Relationships among informal learning environments, teaching procedures and scientific reasoning ability. *International Journal of Science Education*, 23(5), 535-549. <https://doi.org/10.1080/09500690116971>
- Gross, P. R. (2011). *Review of the National Research Council's Framework for K-12 Science Education*. Thomas B. Fordham Institute.
- Han, J. (2013). *Scientific reasoning: Research, development, and assessment*. The Ohio State University (Doctoral dissertation).
- Hardianti, T., & Kuswanto, H. (2017). Difference among levels of inquiry: Process skills improvement at senior high school in Indonesia. *International Journal of Instruction*, 10(2), 119-130. <https://doi.org/10.12973/iji.2017.1028a>
- Harrison, G. M., & Vallin, L. M. (2018). Evaluating the metacognitive awareness inventory using empirical factor-structure evidence. *Metacognition and Learning*, 13(1), 15-38. <https://doi.org/10.1007/s11409-017-9176-z>
- Hawkins, J., & Pea, R. D. (1987). Tools for bridging the cultures of everyday and scientific thinking. *Journal of Research in Science Teaching*, 24(4), 291-307. <https://doi.org/10.1002/tea.3660240404>
- Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education*, 28(1), 1-35. <https://doi.org/10.1007/s10212-011-0097-y>
- Hwang, H., Ha, M., Park, E. (2021). Exploring the effects of overconfidence bias and hard-easy effect in self-monitoring of biological concept test. *Brain, Digital, & Learning*, 11(2), 307-319. <http://doi.org/10.31216/BDL.20210020>
- Inhelder, B., Parsons, A., Milgram, S., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures*. Basic Books.
- Jansen, M., Schroeders, U., & Lüdtke, O. (2014). Academic self-concept in science: Multidimensionality, relations to achievement measures, and gender differences. *Learning and Individual Differences*, 30, 11-21. <https://doi.org/10.1016/j.lindif.2013.12.003>
- Jensen, J. L., Neeley, S., Hatch, J. B., & Piorczynski, T. (2017). Learning scientific reasoning skills may be key to retention in science, technology, engineering, and mathematics. *Journal of College Student Retention: Research, Theory & Practice*, 19(2), 126-144. <https://doi.org/10.1177/1521025115611616>
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- Kalinowski, S. T., & Willoughby, S. (2019). Development and validation of a scientific (formal) reasoning test for college students. *Journal of Research in Science Teaching*, 56(9), 1269-1284. <https://doi.org/10.1002/tea.21555>
- Ku, K. Y., & Ho, I. T. (2010). Metacognitive strategies that enhance critical thinking. *Metacognition and Learning*, 5(3), 251-267. <https://doi.org/10.1007/s11409-010-9060-6>
- Kessels, U., Heyder, A., Latsch, M., & Hannover, B. (2014). How gender differences in academic engagement relate to students' gender identity. *Educational Research*, 56(2), 220-229. <https://doi.org/10.1080/00131881.2014.898916>
- Kind, P. M., & Osborne, J. (2017). Styles of scientific reasoning: A cultural rationale for science education? *Science Education*, 101(1), 8-31. <https://doi.org/10.1002/sce.21251>
- Klahr, D., Zimmerman, C., & Jirout, J. (2011). Educational interventions to advance children's scientific thinking. *Science*, 333(6045), 971-974. <https://doi.org/10.1126/science.1204528>
- Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and individual differences*, 17(2), 161-173. <https://doi.org/10.1016/j.lindif.2007.03.004>



- Kuhn, D. (2002). *What is scientific thinking and how does it develop?* In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development* (pp. 371–393). Blackwell.
- Kuhn, D. (2007). Reasoning about multiple variables: Control of variables is not the only challenge. *Science Education*, 91(5), 710-726. <https://doi.org/10.1002/sce.20214>
- Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, 96(4), 674. <https://doi.org/10.1037/0033-295X.96.4.674>
- Kuhn, D., Amsel, E., O'Loughlin, M., Schauble, L., Leadbeater, B., & Yotive, W. (1988). *The development of scientific thinking skills*. Academic Press.
- Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction*, 9(4), 285-327.
- Kuhn, D., & Dean, Jr, D. (2004). Connecting scientific reasoning and causal inference. *Journal of Cognition and Development*, 5(2), 261-288. [https://doi.org/10.1207/s15327647jcd0502\\_5](https://doi.org/10.1207/s15327647jcd0502_5)
- Lappan, G. (2000). A vision of learning to teach for the 21st century. *School Science and Mathematics*, 100(6), 319-325. <https://doi.org/10.1111/j.1949-8594.2000.tb17326.x>
- Lawson, A. E. (1995). *Science teaching and the development of thinking*. Watsworth Publishing Company.
- Lawson, A. E. (2000). Classroom test of scientific reasoning. *Journal of Research in Science Teaching*, 15(1), 11-24.
- Lawson, A. E. (2003). *The neurological basis of learning, development and discovery*. Kluwer Academic Publishers.
- Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1), 11-24.
- Lawson, A. E., & Lawson, C. A. (1980). A theory of teaching for conceptual understanding, rational thought, and creativity. In A.E Lawson (Ed.), *The psychology of teaching and thinking for creativity* (pp. 103-148). The Ohio State University.
- Lawson, A. E., Adi, H., & Karplus, R. (1979). Development of correlational reasoning in secondary schools: Do biology courses make a difference? *The American Biology Teacher*, 41(7), 420-430. <https://doi.org/10.2307/4446678>
- Lawson, A. E., Banks, D. L., & Logvin, M. (2007). Self-efficacy, reasoning ability, and achievement in college biology. *Journal of Research in Science Teaching*, 44(5), 706-724. <https://doi.org/10.1002/tea.20172>
- Leatham, K. R. (2012). Problems identifying independent and dependent variables. *School Science and Mathematics*, 112(6), 349-358. <https://doi.org/10.1111/j.1949-8594.2012.00155.x>
- Liben, L. S., & Golbeck, S. L. (1980). Sex differences in performance on Piagetian spatial tasks: Differences in competence or performance? *Child Development*, 51, 594-597. <https://doi.org/10.2307/1129301>
- Longeot, F. (1965). Analyse statistique de trois tests genetique collectifs. [Statistical analysis of three collective genetic tests]. *Bulletin de l'Institut National D' Etude*, 20, 219-237.
- Mari, J. S. (2012). Gender related differences in acquisition of formal reasoning schemata: Pedagogic implication of teaching chemistry using process-based approaches. *International Journal for Cross-Disciplinary Subjects in Education (IJCDSE)*, 2(2), 993-997. <https://doi.org/10.20533/ijcdse.2042.6364.2012.0141>
- Martin, M. O., Mullis, I. V. S., Foy, P., & Hooper, M. (2016). *TIMSS 2015 International results in science*. Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2015/international-results/>
- Mingo, M. A., Chang, H., & Williams, R. L. (2018). Undergraduate students' preferences for constructed versus multiple-choice assessment of learning. *Innovative Higher Education*, 43, 143–152. <https://doi.org/10.1007/s10755-017-9414-y>
- Murphy, R. A., Schmeer, S., Vallée-Tourangeau, F., Mondragon, E., & Hilton, D. (2011). Making the illusory correlation effect appear and then disappear: The effects of increased learning. *Quarterly Journal of Experimental Psychology*, 64(1), 24-40. <http://dx.doi.org/10.1080/17470218.2010.493615>
- Neimark, E. D. (1975). Intellectual development during adolescence. *Review of Child Development Research*, 4, 541-594.
- Niaz, M., & Robinson, W. R. (1992). From 'algorithmic mode' to conceptual gestalt in understanding the behaviour of gases: An epistemological perspective. *Research in Science and Technological Education*, 10(1), 53-64. <https://doi.org/10.1080/0263514920100105>
- Nieminen, P., Savinainen, A., & Viiri, J. (2012). Relations between representational consistency, conceptual understanding of the force concept, and scientific reasoning. *Physical Review Special Topics - Physics Education Research*, 8(1), Article 010123. <https://doi.org/10.1103/PhysRevSTPER.8.010123>
- Novia, N., & Riandi, R. (2017). The analysis of students scientific reasoning ability in solving the modified Lawson Classroom Test of scientific reasoning (MLCTSR) problems by applying the levels of inquiry. *Jurnal Pendidikan IPA Indonesia*, 6(1). <https://doi.org/10.15294/jpii.v6i1.9600>
- NRC [National Research Council]. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: National Academy Press.
- OECD. (2018). *Preparing our Youth for an Inclusive and Sustainable World: The OECD PISA Global Competence Framework*. OECD Library.
- Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, 48(1), 341-348. <https://doi.org/10.3758/s13428-015-0576-1>
- Piaget, J. (1965). *The moral development*. Free Press.
- Piraksa, C., Srisawasdi, N., & Koul, R. (2014). Effect of gender on student's scientific reasoning ability: A case study in Thailand. *Procedia-Social and Behavioral Sciences*, 116, 486-491. <https://doi.org/10.1016/j.sbspro.2014.01.245>
- Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, 146(3642), 347-353.
- Pohl, R., & Pohl, R. F. (Eds.). (2004). *Cognitive Illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press.





- Raven, R. J. (1973). The development of a test of Piaget's logical operations. *Science Education*, 57, 33-40.
- Roadrangka, V., Yeany, R. H., & Padilla, M. J. (1983). The construction and validation of group assessment of logical thinking (GALT). In Paper presented at the annual meeting of the National Association for Research in Science Teaching, Dallas, TX.
- Rogers, P., Hattersley, M., & French, C. C. (2019). Gender role orientation, thinking style preference and facets of adult paranormality: A mediation analysis. *Consciousness and Cognition*, 76, Article 102821. <https://doi.org/10.1016/j.concog.2019.102821>
- Rusmana, A. N., Roshayanti, F., & Ha, M. (2020). Debiasing overconfidence among Indonesian undergraduate students in the biology classroom: An intervention study of the KAAR model. *Asia-Pacific Science Education*, 6(1), 228-254. <https://doi.org/10.1163/23641177-BJA00001>
- Rusmana, A. N., Sya'bandari, Y. Aini, R. Q., Rachmatullah, A., & Ha, M. (2021). Teaching Korean science for Indonesian middle school students: Promoting Indonesian students' attitude towards science through the global science exchange programme. *International Journal of Science Education*, 43(11), 1837-1859. <https://doi.org/10.1080/09500693.2021.1938278>
- Rustaman. (2009). *Analisis Konten dan Capaian Sains Siswa Indonesia dalam TIMSS [Trends in International Mathematics and Science Study] tahun 1999, 2003, dan 2007*. Badan Penelitian Pengembangan Departemen Pendidikan Nasional.
- Saad, M. I. M., Baharom, S., & Mokhsein, S. E. (2017). Scientific reasoning skills based on socio-scientific issues in the biology subject. *International Journal of Advanced and Applied Sciences*, 4(3), 13-18. <https://doi.org/10.21833/ijaas.2017.03.003>
- Severiens, S. E., & ten Dam, G. T. M. (1997). Gender and gender identity differences in learning styles. *Educational Psychology*, 17, 79-93. <https://doi.org/10.1080/0144341970170105>
- Shayer, M., & Wharry, D. (1975). *Piaget in the classroom: I. Testing a whole class at the same time*. Chelsea College University of London.
- She, H. C., & Liao, Y. W. (2010). Bridging scientific reasoning and conceptual change through adaptive web-based learning. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, 47(1), 91-119. <https://doi.org/10.1002/tea.20309>
- Sladek, R. M., Bond, M. J., & Phillips, P. A. (2010). Age and gender differences in preferences for rational and experiential thinking. *Personality and Individual Differences*, 49(8), 907-911. <https://doi.org/10.1016/j.paid.2010.07.028>
- Soyibo, K. (1999). Gender differences in Caribbean students' performance on a test of errors in biological labeling. *Research in Science and Technological Education*, 17(1), 75-82.
- Spelke, E. S. (2005). Sex differences in intrinsic aptitude for mathematics and science? A critical review. *American Psychologist*, 60(9), 950-958. <https://doi.org/10.1037/0003-066X.60.9.950>
- Star, C., & Hammer, S. (2008). Teaching generic skills: Eroding the higher purpose of universities, or an opportunity for renewal? *Oxford Review of Education*, 34(2), 237-251. <https://doi.org/10.1080/03054980701672232>
- Suyanto, S. (2018). The implementation of the scientific approach through 5Ms of the revised curriculum 2013 in Indonesia. *Cakrawala Pendidikan*, 37(1), 22-29. <https://doi.org/10.21831/cp.v37i1.18719>
- Sya'bandari, Y., Ha, M., Lee, J. K., & Shin, S. (2019). The relation of gender and track on high school students' attitude toward convergence. *Journal of Baltic Science Education*, 18(3), 417-434. <https://dx.doi.org/10.33225/jbse/19.18.417>
- Tobin, K. G., & Capie, W. (1980). Teaching process skills in the middle school. *School Science and Mathematics*, 80, 590-600. <https://doi.org/10.1111/j.1949-8594.1980.tb09745.x>
- Tobin, K. G., & Capie, W. (1981). The development and validation of a group test of logical thinking. *Educational and Psychological Measurement*, 41(2), 413-423. <http://dx.doi.org/10.1177/001316448104100220>
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39(7), 1275. <https://doi.org/10.3758/s13421-011-0104-1>
- Tozoglu, D., Tozoglu, M. D., Gurses, A., & Dogar, C. (2004). The students' perceptions: Essay versus multiple-choice type exams. *Journal of Baltic Science Education*, 3(2), 52-59. <http://www.scientiasocialis.lt/jbse/?q=node/77>
- UNESCO. (2017). *Cracking the code: Girls' and women's education in science, technology, engineering and mathematics (STEM)*. The United Nations Educational, Scientific and Cultural Organization.
- Utomo, A. P., Narulita, E., & Shimizu, K. (2018). Diversification of reasoning science test items of TIMSS grade 8 based on higher order thinking skills: A case study of Indonesian students. *Journal of Baltic Science Education*, 17(1), 152-161. <https://dx.doi.org/10.33225/jbse/18.17.152>
- Valanides, N. C. (1996). Formal reasoning and science teaching. *School Science and Mathematics*, 96(2), 99-107.
- Valanides, N. (1997). Formal reasoning abilities and school achievement. *Studies in Educational Evaluation*, 23(2), 169-185. [https://doi.org/10.1016/S0191-491X\(97\)00011-4](https://doi.org/10.1016/S0191-491X(97)00011-4)
- Van der Graaf, J., Van de Sande, E., Gijssels, M., & Segers, E. (2019). A combined approach to strengthen children's scientific thinking: Direct instruction on scientific reasoning and training of teacher's verbal support. *International Journal of Science Education*, 41(9), 1119-1138. <http://doi.org/10.1080/09500693.2019.1594442>
- Vass, E., Schiller, D., & Nappi, A. J. (2000). The effects of instructional intervention on improving proportional, probabilistic, and correlational reasoning skills among undergraduate education majors. *Journal of Research in Science Teaching*, 37(9), 981-995. [https://doi.org/10.1002/1098-2736\(200011\)37:9%3C981::AID-TEA7%3E3.0.CO;2-1](https://doi.org/10.1002/1098-2736(200011)37:9%3C981::AID-TEA7%3E3.0.CO;2-1)
- Ward, S. J., & King, L. A. (2018). Gender differences in emotion explain women's lower immoral intentions and harsher moral condemnation. *Personality and Social Psychology Bulletin*, 44(5), 653-669. <https://doi.org/10.1177/0146167217744525>
- Ward, S. J., & King, L. A. (2020). Examining the roles of intuition and gender in magical beliefs. *Journal of Research in Personality*, 86, 103956. <https://doi.org/10.1016/j.jrp.2020.103956>
- Warren, J. R. (1979). Some Specifics in General Education. Paper presented at the Annual Metropolitan Conference on General Education and Entering Learners, New Jersey.
- Wasis. (2014). Analyzing physics items of UN, TIMSS, and PISA-based on higher-order thinking and scientific literacy. In Sutrisno,



H., Dwandaru, W. S. B., Krisnawan, K. P., Darmawan, D., Priyambodo, E., Yulianty, & E. Nurohmah, S. (Eds.), *Proceeding of international conference on research, implementation and education of mathematics and sciences 2014* (pp. 147-154). Yogyakarta State University.

Weld, J., Stier, M., & McNew-Birren, J. (2011). The development of a novel measure of scientific reasoning growth among college freshmen: The constructive inquiry science reasoning skills test. *Journal of College Science Teaching*, 40(4), 101-107.

Wilhelm, J., Cole, M., Cohen, C., & Lindell, R. (2018). How middle level science teachers visualize and translate motion, scale, and geometric space of the Earth-Moon-Sun system with their students. *Physical Review Physics Education Research*, 14(1), Article 010150. <https://doi.org/10.1103/PhysRevPhysEducRes.14.010150>

Woolley, J. S., Deal, A. M., Green, J., HATHENBRUCK, F., Kurtz, S. A., Park, T. K., Pollock, S. V., Transtrum, M. B., & Jensen, J. L. (2018). Undergraduate students demonstrate common false scientific reasoning strategies. *Thinking Skills and Creativity*, 27, 101-113. <https://doi.org/10.1016/j.tsc.2017.12.004>

Yang, F. (2004). Exploring high school students' use of theory and evidence in an everyday context: The role of scientific thinking in environmental science decision-making. *Journal of Science Education*, 26(11), 1345-1364. <https://doi.org/10.1080/0950069042000205404>

Yang, F. Y., Bhagat, K. K., & Cheng, C. H. (2019). Associations of epistemic beliefs in science and scientific reasoning in university students from Taiwan and India. *International Journal of Science Education*, 41(10), 1347-1365. <https://doi.org/10.1080/09500693.2019.1606960>

Yang, I. H., Kwon, Y. J., Kim, Y. S., Jang, M. D., Jeong, J. W., & Park, K. T. (2002). Effects of students' prior knowledge on scientific reasoning in density. *Journal of the Korean Association for Science Education*, 22(2), 314-335.


Yanto, B. E., Subali, B., & Suyanto, S. (2019). Improving students' scientific reasoning skills through the three levels of inquiry. *International Journal of Instruction*, 12(4), 689-704. <http://dx.doi.org/10.29333/iji.2019.12444a>

Zhou, S., Han, J., Koenig, K., Raplinger, A., Pi, Y., Li, D., ... & Bao, L. (2016). Assessment of scientific reasoning: The effects of task context, data, and design on student reasoning in control of variables. *Thinking Skills and Creativity*, 19, 175-187. <https://doi.org/10.1016/j.tsc.2015.11.004>

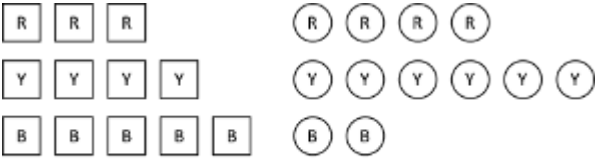
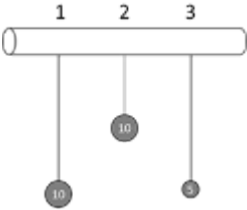
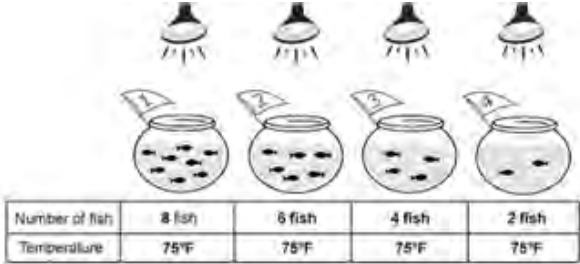
**Appendixes**

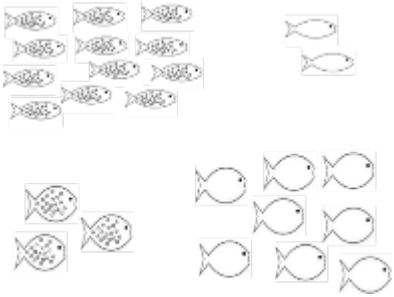
**Appendix 1.**

*The Items with gap score*

Item	Questions	Gap score
Gap between the USA and Indonesia university students		
1	<p>Hannah wants to know which type of soil is best for growing corn. She also wants to know which type of fertilizer is best. She performs an experiment using two types of soil (A and B) and two types of fertilizer (X and Y). The figure below shows what her corn looks like at the end of the summer:</p> <p style="text-align: center;">Same amount of water and same amount of light</p> <div style="text-align: center;">  </div> <p>What can Hannah conclude from this experiment?</p> <p>a) Soil B is best for growing her corn                      b) Fertilizer Y is best for growing her corn                      c) Soil B is best for growing her corn, and Fertilizer Y is best for growing her corn                      d) It is not possible to conclude which soil or which fertilizer is best for growing her corn*</p>	48%




Item	Questions	Gap score										
8	<p>Twelve wooden blocks and twelve wooden balls are placed in a bag (see diagram). Some are red (R), some are yellow (Y), and some are blue (B). Someone reaches into the bag and pulls out an object without looking at the color or feeling the shape. What is the chance the object is a red or blue ball?</p>  <p style="text-align: center;">12 blocks and 12 balls</p> <p>What is the chance the object is a red or blue ball?</p> <p>a) 1 chance out of 2 b) 1 chance out of 3 c) 1 chance out of 4* d) 1 chance out of 6 e) None of the above</p>	38%										
Gap between Indonesia university and high school students												
7	<p>The drawing on the right shows three strings hanging from a bar. Each string has a metal weight at the end that weighs 5 or 10 ounces. The weights can be swung back and forth, and the time it takes for the weight to swing back and forth can be measured. Suppose you want to find out whether the length of the string has an effect on how long it takes for the string to swing back and forth.</p>  <p>Which string(s) would you use to find out?</p> <p>a) Any string                      d) 1 and 3 b) All 3 strings                  e) 2 and 3 c) 1 and 2*</p>	30%										
20	<p>A student is interested in the behavior of fish. He puts 8 fish in a bowl, 6 fish in a second bowl, 4 fish in a third bowl, and 2 fish in a fourth bowl. He places each fish bowl under light and he keeps the temperature at 75°F for all four bowls. What can the student learn about fish from doing just this experiment?</p>  <table border="1" data-bbox="409 1705 988 1771"> <tbody> <tr> <td>Number of fish</td> <td>8 fish</td> <td>6 fish</td> <td>4 fish</td> <td>2 fish</td> </tr> <tr> <td>Temperature</td> <td>75°F</td> <td>75°F</td> <td>75°F</td> <td>75°F</td> </tr> </tbody> </table> <p>a) If the number of fish in the fish bowl affects the behavior of the fish* b) If the temperature of the fish bowl affects the behavior of the fish c) If the temperature of the fish bowl and the amount of light affect the behavior of the fish d) If the number of fish, the temperature, and the amount of light affect the behavior of the fish</p>	Number of fish	8 fish	6 fish	4 fish	2 fish	Temperature	75°F	75°F	75°F	75°F	25%
Number of fish	8 fish	6 fish	4 fish	2 fish								
Temperature	75°F	75°F	75°F	75°F								

Item	Questions	Gap score
Gap between male and female Indonesian students		
10	<p>Katherine catches 25 fish (see diagram). The fish are all same species but vary in shape and whether they have spots.</p>  <p>Does there seem to be a relationship between the shape of the fish and whether it has spots?</p> <p>a) There seems to be a relationship* b) There seems to be no relationship c) You can't tell from this sample</p>	7%
4	<p>A large number of sea gulls died and washed up on a beach. A biologist wants to know why. She finds that most of the dead gulls have plastic trash in their stomach, and this makes her wonder whether plastic trash is harming the birds. What future research would be most useful to find out?</p> <p>a) Search for more dead sea gulls on the beach, and determine if they have plastic in their stomachs b) Search for sea gulls that look weak or sick, and perform an ultrasound scan to see if they have plastic in their stomachs c) Catch sea gulls that look healthy, and perform an ultrasound scan to see if they have plastic in their stomachs*</p>	7%

Note : \* = the correct answer

**Appendix 2**

*The Items with distractor believed as the correct answer*

Item	Questions	Correct answer	The most chosen distractor
1	<p>Hannah wants to know which type of soil is best for growing corn. She also wants to know which type of fertilizer is best. She performs an experiment using two types of soil (A and B) and two types of fertilizer (X and Y). The figure below shows what her corn looks like at the end of the summer:</p> <p style="text-align: center;">Same amount of water and same amount of light</p>  <p>What can Hannah conclude from this experiment?</p> <p>a. Soil B is best for growing her corn. b. Fertilizer Y is best for growing her corn. c. Soil B is best for growing her corn, and Fertilizer Y is best for growing her corn. d. It is not possible to conclude which soil or which fertilizer is best for growing her corn.</p>	D	C



Item	Questions	Correct answer	The most chosen distractor															
13	<p>A medical researcher wants to know if high density lipoproteins (HDLs) contribute to high blood pressure in overweight men. She tests for HDLs in 240 overweight male patients and measures their blood pressure. The table at left shows the number of men with HDLs and without HDLs that have high blood pressure or not. Does there appear to be an association between HDLs and high blood pressure?</p> <table style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th colspan="2"></th> <th colspan="2">High blood pressure</th> </tr> <tr> <th colspan="2"></th> <th>Yes</th> <th>No</th> </tr> </thead> <tbody> <tr> <th rowspan="2">HDLs present</th> <th>HDLs present</th> <td style="border: 1px solid black;">150</td> <td style="border: 1px solid black;">50</td> </tr> <tr> <th>HDLs absent</th> <td style="border: 1px solid black;">30</td> <td style="border: 1px solid black;">10</td> </tr> </tbody> </table> <p>a. Yes. b. No. c. You can't tell from this data.</p>			High blood pressure				Yes	No	HDLs present	HDLs present	150	50	HDLs absent	30	10	B	A
		High blood pressure																
		Yes	No															
HDLs present	HDLs present	150	50															
	HDLs absent	30	10															

Received: September 14, 2021

Accepted: December 04, 2021

Cite as: Ha, M., Sya'bandari, Y., Rusmana, A. N., Aini, R. Q., & Fadillah, S. M. (2021). Comprehensive analysis of the FORT instrument: Using distractor analysis to explore students' scientific reasoning based on academic level and gender difference. *Journal of Baltic Science Education*, 20(6), 906-926. <https://doi.org/10.33225/jbse/21.20.906>

**Minsu Ha**  
(Corresponding author)

PhD, Professor, Department of Science Education, College of Education, Kangwon National University, 1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do, 24341 Republic of Korea.  
E-mail: msha@kangwon.ac.kr  
ORCID: <https://orcid.org/0000-0003-3087-3833>

**Yustika Sya'bandari**

MEd, Research Assistant, Department of Science Education, College of Education, Kangwon National University, 1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do, 24341 Republic of Korea.  
E-mail: yustikasya@gmail.com  
ORCID: <https://orcid.org/0000-0002-3432-1203>

**Ai Nurlaelasari Rusmana**

MEd, Research Assistant Department of Science Education, College of Education, Kangwon National University, 1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do, 24341 Republic of Korea.  
E-mail: ainurlaelasarirusmana@gmail.com  
ORCID: <https://orcid.org/0000-0003-1281-168X>

**Rahmi Qurota Aini**

MEd, Doctoral Student, Department of Biology, Middle Tennessee State University, Murfreesboro, Tennessee, the United States.  
Email: rqa2a@mtmail.msu.edu

**Sarah Meilani Fadillah**

SPd, Master Student, Department of Science Education, College of Education, Kangwon National University, 1 Kangwondaehak-gil, Chuncheon-si, Gangwon-do, 24341 Republic of Korea.  
E-mail: sarahmeilani.sm@gmail.com

