MALAYSIAN JOURNAL OF LEARNING AND INSTRUCTION
https://e-journal.uum.edu.my/index.php/mjli

How to cite this article:
Sarah Meilani Fadillah, Minsu Ha, Eni Nuraeni & Nurma Yunita Indriyanti. (2023). Exploring confidence accuracy and item difficulty in changing multiple-choice answers of scientific reasoning test. *Malaysian Journal of Learning and Instruction, 20*(2), 319-341. https://doi.org/10.32890/mjli2023.20.2.5

EXPLORING CONFIDENCE ACCURACY AND ITEM DIFFICULTY IN CHANGING MULTIPLE-CHOICE ANSWERS OF SCIENTIFIC REASONING TEST

[1]Sarah Meilani Fadillah, [2]Minsu Ha, [3]Eni Nuraeni & [4]Nurma Yunita Indriyanti
[1]Department of Science Education,
Kangwon National University, Chuncheon, South Korea
[2]Department of Biology Education,
Seoul National University, Seoul, South Korea
[3]Department of Biology Education,
Universitas Pendidikan Indonesia, Bandung, Indonesia
[4]Department of Science Education,
Sebelas Maret University, Surakarta, Indonesia

[2]*Corresponding author: msha101@snu.ac.kr*

Received: 13/10/2023 Revised: 18/5/2023 Accepted: 28/5/2023 Published: 31/7/2023

ABSTRACT

Purpose – Researchers discovered that when students were given the opportunity to change their answers, a majority changed their responses from incorrect to correct, and this change often increased the overall test results. What prompts students to modify their answers? This study aims to examine the modification of scientific

reasoning test, with additional exploration on confidence accuracy and its relation to item difficulty.

Methodology – A set of pre-test and post-test experiments which included 20 items of scientific reasoning test with confidence judgement on each item were used. The items of the instruments were assessed for their validity by analysing their psychometric properties using the three-parameter (3PL) Item Response Theory, which was carried out in R studio. The set of items were randomly administered to 205 Indonesian undergraduate students with a background in science education related major. The accuracy of confidence was determined by categorising correct or incorrect answers to scientific reasoning questions based on their level of confidence.

Findings – The results revealed that responses were modified more frequently from incorrect to correct than from correct to incorrect, resulting in a significant gain in overall scientific reasoning score although these modifications were not shown to be connected to the item's difficulty level. Even though confidence level also increased significantly, it was observed that Indonesian students repeatedly responded with overconfidence even after sitting for the same test after three weeks, which could indicate a lack of metacognitive ability. The findings of this study serve to spur educators to begin actively engaging in metacognitive training in their teaching and learning activities as a result of overconfidence that frequently occurs among Indonesian students in examinations.

Significance – This study provides further substantiation in the field of scientific reasoning and cognitive science; that a trend of confidence accuracy change in scientific reasoning test has been observed. It also contributes to uncovering the true ability of Indonesian students when performing such reasoning tests through their repeated attempts.

Keywords: Item difficulty, metacognitive judgement, overconfidence bias, scientific reasoning, confidence accuracy, revision study, undergraduate students.

INTRODUCTION

There is a widely held belief that changing initial answers to objective test questions tends to lower scores (Benjamin et al., 1984; Friedman-

Erickson, 1994). However, a well-known study by Kruger et al. (2005) found that when students were given a chance to change their answers, they typically changed from incorrect to correct answers and most of them usually improved their test scores. Recent research on revisions is now connecting it to metacognition by focusing on how changes in confidence judgements are related, as inspired by Kruger's work. Couchman et al. (2016) explored the grounds for revision and emphasised the difference between a basic correction and a more complex cognitive process that occurs during revision. In this study, the authors advocated the inclusion of a metacognitive approach in the study of revisions. Additionally, metacognition is an important skill because not knowing the extent to which an individual lacks knowledge or skills can lead to the implementation of suboptimal strategies during study, which in turn may hinder learning and performance (Coutinho et al., 2021).

Measuring metacognitive abilities, however, is a difficult task (Craig et al., 2020). This is especially crucial when metacognition is measured by a self-report instrument, which may be erroneous due to overestimation bias. Thus, rather than using only a single self-report measurement of metacognition, another measurement such as accuracy of confidence on scientific reasoning test should also be used as a complement to quantify metacognitive changes. Recently, there has been an increasing interest in investigating metacognitive processes associated with reasoning studies. Metacognitive mechanisms are thought to play a role in scientific reasoning, as demonstrated by research that observed differences in confidence scores when cognitive abilities were held constant (Ackerman & Thompson, 2017; Fritzsche et al., 2012). Accordingly, Kruger and Dunning (1999) argued that to accurately measure self-report assessment, especially metacognitive ability, an individual needs to acquire a good scientific reasoning ability. Prior studies have used content knowledge questions during revision study. For instance, a recent revision study by Merry et al. (2021) utilised questions on human anatomy and physiology, introductory biology, and neuroscience topics. In this study, we utilised a reasoning task instead of content knowledge questions because content knowledge questions might be related to one's ability to memorise.

However, modifications on a test could also interfere with each item's psychometrics properties, as well as with item parameters (Papanastasiou, 2015). In other words, another variable such as item difficulty could be a potential factor that could be linked to the

examination of revisions. This study may provide additional literature for future research on revision that also highlights confidence judgement on scientific reasoning abilities as metacognitive judgement. Most revision studies that use a metacognitive approach come from Western countries (Coutinho et al., 2020). However, in non-Western countries, particularly among Indonesian undergraduate participants, the use of reasoning tasks to demonstrate confidence accuracy in revision studies is rather limited.

Revising Answers in Examinations

Educators generally believe that students should stick to their original answers and avoid changing them during testing situations (Kruger et al., 2005). However, several findings from a study on revisions reported since the 1960s indicate that this widely held belief is factually inaccurate (Archer & Pippert, 1962; Benjamin et al., 1984; Wagner et al., 1998; Papanastasiou, 2015; Stylianou-Georgiou & Papanastasiou, 2017). Three influential psychologists, Kruger, Wirtz, and Miller (2005) conducted thorough reviews of several studies focusing on revision research. They discovered that answer changes from incorrect to correct mostly occurred, and that most people who changed their answers typically improved their test scores. Kruger et al. (2005) coined the term "first instinct fallacy" to refer to the belief that one should always follow one's instincts.

Revising answers has been shown to be beneficial in examinations, as most revisions are from incorrect to correct, thereby increasing scores (Kruger et al., 2005). Ballance (1977) reported that scores of more than half of their participants increased after revision. Although, changing answers does not always result in the right answer, several studies found a higher propensity for wrong-to-right changes compared to right-to-wrong changes. Merry et al. (2021) reported in their revision study that students who changed their answers from wrong to right were more in number as compared to students who changed answers from right to wrong. Stylianou-Georgiou and Papanastasiou (2017) reported that students are more likely to change their answers from wrong-to-right rather than from right-to-wrong changes or wrong-to-wrong changes. The phenomenon of changing minds occurs not only during examinations, but also in general situations. According to Stone et al. (2022) there are two significant factors that can influence the process of changing minds when it comes to value-based decision-making: decision uncertainty and subjective confidence. In the same

review, Stone et al. (2022) argued that when the confidence level is low and the uncertainty is high, people are more likely to change their minds, and their efficacy is primarily determined by metacognitive sensitivity.

Confidence Accuracy and Item Difficulty Linked to Modifications on Scientific Reasoning Test

In 1979, Flavell came up with the concept of metacognition as thinking about one's own thoughts. Since then, researchers on metacognitive calibration have adopted various strategies to identify the actual metacognitive performance (Bol & Hacker, 2001). Previous studies have discussed much on the relationship between metacognition and reasoning. Kruger and Dunning (1999) argued that the ability of individuals to monitor their own capacity needs a high level of analytical thinking capacity. To determine whether an individual's skill is adequate, they must first be proficient in thinking analytically; otherwise, they remain unaware of their capacity. In measuring metacognition, confidence accuracy on cognitive tasks and self-report instruments are frequently used by cognitive scientists. However, Craig et al. (2020) argued that identifying metacognition through self-reporting measures might be problematic because an individual might not make judgement accurately. Kruger and Dunning (1999) discovered that low performers are more likely to evaluate their capacity to be just above average; on the other hand, students in the top quartile tend to underestimate their ability.

In the literature, researchers also use the accuracy of confidence on reasoning tasks to calculate metacognition. Pieschl (2009) suggests that the ability to monitor errors is related to metacognitive ability whereby students give self-judgement on their answers. Further, the method was specifically termed as calibration. Schraw et al. (2013), and subsequently Rutherford (2017), investigated the usage of local judgements that resulted in a 2×2 data matrix model that represents associations between variables. The association between variables include: correct answer and confident; correct answer and not confident; incorrect answer and confident; and incorrect answer and not confident. Further, revision studies added item difficulty correlation to answer modification.

Couchman et al. (2016) suggested adding metacognitive measurement to the study of revisions. They argued that the approach used by

Kruger et al. (2005) did not provide reasons behind revisions, did not indicate the difference between simple mis-markings and cognitively laborious re-thinking, and was unable to determine which type of decision(s) was most beneficial, i.e., revising or not revising. Thus, Couchman et al. (2016) proposed to add metacognitive measures to clarify on the psychology behind it; arguing that metacognitive judgements hold importance in predicting accuracy in exam situations and could determine whether they prove to be a useful guide for marking revisions.

After Couchman et al. (2016), subsequent revision studies such as by Kruger et al. (2005) began to add metacognitive perspectives in their revision studies. Among them was a study by Stylianou-Georgiou and Papanastasiou (2017) who observed that confidence was negatively associated with unsuccessful answer changing, with students being less likely to make unsuccessful answer changes on items that were responded to more confidently. Stylianou-Georgiou and Papanastasiou (2017) suggested that if confidence judgement is informed appropriately, errors due to question misinterpretation or misreading are more likely to be monitored well. Additionally, they also found a significant association between item difficulty and unsuccessful answer modifications made by students. Changes made to a test have the potential to interfere with the psychometric properties and item parameters of each test item (Papanastasiou, 2015). To the best of our knowledge, Stylaniou-Georgiou and Papanastasiou (2017) were the first to examine item difficulty association with answer modification. Students tend to focus on items that are not familiar to them (Metcalfe & Kornell, 2005) and it was found that students were frequently changing answers to difficult questions (Stylaniou-Georgiou & Papanastasiou, 2017).

Kalinowski and Willoughby (2019) have developed and validated a scientific reasoning test using the three parameters (3PL) Item Response Theory (IRT) model. By implementing the 3PL-IRT model, item psychometrics including item difficulty can be acknowledged. While May and Jackson (2005) explored various combinations of 3PL-IRT to examine the construct validity of pretest and posttest and its gain scores. Besides, a clearer description of the connection between item parameters and the reliability and validity of items and gain scores could lend insight into methods for enhancing measurement precision (May & Jackson, 2005). However, in this study, we would

like to use the 3PL-IRT model to explore the pattern of answer changes in reasoning tests. According to Ha et al. (2021), there is an emerging issue in terms of empowering reasoning capacity among Indonesian students. By giving a second chance for the students to redo the reasoning task, we are eager to explore whether students are getting more deliberate and thoughtful by monitoring more accurately their answers for the second chance.

Research Questions

The purpose of this study is to investigate how students change their answers in a general reasoning test, by examining how their level of confidence changes as well as identifying how accurate their confidence is in relation to the difficulty of the questions. The research questions of this study are as follows:

1. What are the changes in the general reasoning test in terms of confidence level and Metacognitive Analogy Instruction (MAI) response between the pretest and the posttest?
2. What is the variation in the confidence accuracy change from the pretest to the posttest?
3. What are modifications of answers related to the difficulty of test items?

METHODOLOGY

Participants

This study was carried out at the science education faculty of two public universities in Indonesia. The selection of these universities was based on their willingness to participate and the approval of their faculty members. A total of 231 students who volunteered to take part in the study were chosen randomly. The students from both the universities were registered in science related majors consisting of science education, biology education, and chemistry education. Participants with missing data were excluded from the analysis. With regard to the participants, 116 (52.7%) students came from a university in Central Java Province, while 104 (47.3%) students came from a university in West Java Province. Thus, only 205 students

were able to fully complete the pretest and posttest. Based on the 205 participants, 8.78 percent were male and 91.22 percent were female.

Instruments

In this study, we examined the confidence accuracy changes which occurred when Indonesian students completed a scientific reasoning test twice. A total of 20 items from the FORT instrument (Kalinowski & Willoughby, 2019) were used to assess students' scientific reasoning ability in the pretest and posttest. The FORT instrument had previously been translated into Indonesian by three experts in science education, with Indonesian as their first language. They were also involved in the translation process of the study to ensure readability and content clarity of the instrument. Besides, this instrument was administered to Indonesian participants by Ha et al. (2021). The students' confidence level was determined for each reasoning item using a 5-point Likert scale (ranging from strongly not confident to strongly confident). Additionally, the participants completed the Metacognitive Awareness Instrument (MAI) (Harrison & Vallin, 2018), which consists of 19 items. The MAI assesses students' metacognition within the context of knowledge and regulation. FORT has been used in previous research on scientific reasoning on Indonesian students (Ha et al., 2021). This study adapted the original instrument of FORT to repeat the validity test to ensure the instrument's validity for this study (Kalinowski & Willoughby, 2019). To accomplish this, Messick's framework (1995) was used by performing IRT-Rasch, which determined the item and personal reliability that were run from R studio using TAM package. IRT is a set of statistical methods that can predict the ability of students in completing a certain test. In the three-parameter (3PL) Item Response Theory, there are three parameters that could determine the ability of students in completing the test which are: item difficulty, item discrimination, and the guessing rate. The item's quality was analysed using weighted mean square (MNSQ) equivalent of the infit MNSQ and the unweighted mean square (MNSQ) equivalent of the outfit MNSQ. MNSQs that are acceptable typically vary between 0.5 and 1.5 logits (Wright & Linacre, 1994). Along with the reliability analysis results, we included Cronbach's alpha value to evaluate the internal consistency. MNSQ was found to be within the range for all FORT items (infit= .94 – 1.12, outfit= .92–1.21). In terms of validity, the item reliability, person reliability, and Cronbach's alpha value were found to be .55, .50, and .55, respectively. Validity evidence including

item difficulty and person ability for each item were determined by utilising the three-parameter (3PL) Item Response Theory (Table 1).

Table 1

Psychometric Properties of FORT Instrument

Reasoning Item	Item Difficulty	Item Discrimination	Guessing Rate
Item 1	1.55	1.74	0.05
Item 2	1.55	0.66	0.05
Item 3	0.66	1.38	0.05
Item 4	0.79	0.38	0.06
Item 5	0.46	0.74	0.05
Item 6	0.30	1.02	0.07
Item 7	-0.37	1.18	0.05
Item 8	0.98	0.88	0.05
Item 9	2.33	-0.40	0.05
Item 10	1.53	-0.63	0.06
Item 11	0.17	1.88	0.05
Item 12	1.94	0.57	0.06
Item 13	3.13	1.32	0.05
Item 14	2.46	1.90	0.08
Item 15	1.83	-0.74	0.05
Item 16	1.28	0.75	0.05
Item 17	-0.84	1.22	0.06
Item 18	1.26	0.78	0.06
Item 19	1.45	1.26	0.05
Item 20	-0.02	1.13	0.04

Data Collection

The set of instruments for data collection was administered to the participants through Google survey. Given the limited emphasis on the intervention's effect, the primary objective of this study was to concentrate on the patterns of modification. A one-group pretest-posttest (Cohen et al., 2002) was employed, wherein all participants sat for both, the pretest and posttest. Two weeks after the pretest, the students received a feedback email that included information about their previous pretest responses but did not include information on the correctness or incorrectness of their initial responses. Before taking the posttest, the students were verbally informed of the total pretest

results, which indicated that their scores were low. They were asked to retake the same test with greater attention. The posttest in this study was conducted in the third week after the pretest. Usually, if the test is repeated between the third and sixth week after the pretest, the students are considered not to have learned much, nor remembered how they answered the first time they sat for the test (Brown et al., 2008).

Data Analysis

This study aimed to examine a confidence judgement change on scientific reasoning test in a revision study and its relation to item difficulty. The changes between the pretest and posttest was accomplished by conducting paired t-test analyses on variables such as scientific reasoning score, confidence rating scale, metacognitive self-report scale (MAI), and calibration measurement. In this study, we calculated the calibration score based on both the accuracy of scientific reasoning test and the level of confidence expressed by the participants. The confidence level was rated on a 5-point Likert scale, where 1 (strongly not confident), 2 (not confident), 3 (slightly confident), 4 (confident), and 5 (strongly confident). In this study, the first two points were considered indicative of low confidence and the last three points were considered as indicative of high confidence. The following framework has been previously implemented in a metacognitive judgement study (Lundeberg & Mohan, 2009).

The correct or incorrect answers to the items of the scientific reasoning test are categorised based on their level of confidence. Right answers with a confidence score of one to two are considered correct or right answers but with low confidence, are referred to as RLC (right, low confidence). Right answers with a confidence score of three to five are considered correct and with high confidence, are referred to as RHC (right, high confidence). The same level of confidence is also applied to categorise incorrect answers. Answers that are incorrect or wrong but with low confidence are referred to as WLC (wrong, low confidence), while those with high confidence are called WHC (wrong, high confidence). We calculated the overall confidence judgement change from the pretest to the posttest, and the most frequently occurring change(s) in each item. For the item difficulty of reasoning test, we used three-parameter logistic (3PL) based on the Item Response Theory (IRT) model as previously it was also conducted in

the original version of FORT instrument (Kalinowski & Willoughby, 2019). The item difficulty value indicates the ability of students and how difficult it is to correctly answer the item in comparison to other questions. A value below zero indicates that the item is easier than average, whereas an item with a positive value indicates that the item is difficult (Kalinowski & Willoughby, 2019). We then calculated the correlation between the pretest-to-posttest changes to item difficulty.

RESULTS

Change in General Reasoning, Confidence, and Confidence Accuracy

To examine metacognitive judgement, and the change among the Indonesian students, this study utilised paired t-test analysis between the pretest and postest of the reasoning, confidence, and MAI variables. We also compared the calibration of reasoning and confidence. The overall change between the pretest and the posttest are shown in Table 2 as follows.

Table 2

Overview of Reasoning, Confidence, MAI, and Calibration Results

		Mean	*SD*	*t*-value	*p*-value
Reasoning	Pre	0.33	0.14	-3.75	0.00
	Post	0.36	0.15		
Confidence	Pre	3.71	0.45	-2.43	0.02
	Post	3.79	0.56		
MAI	Pre	3.89	0.47	1.03	0.30
	Post	3.87	0.50		
RHC	Pre	0.31	0.14	-3.90	0.00
	Post	0.34	0.15		
WLC	Pre	0.09	0.10	4.33	0.00
	Post	0.06	0.09		
RLC	Pre	0.02	0.04	0.00	1.00
	Post	0.02	0.04		
WHC	Pre	0.58	0.16	0.14	0.89
	Post	0.58	0.17		

Firstly, as shown in Table 2, the reasoning score, as measured by FORT, showed a significant increase from pretest (M_{pre}= .33) to posttest (M_{post}= .36, p<.01). This suggests that the students' ability to reason scientifically improved in the second chance of completing an identical scientific reasoning test. Secondly, the confidence score, which reflects students' self-assessment in their answers, also exhibited a significant increase from pretest (M_{pre}= 3.71) to posttest (M_{post}= 3.79, p=.02). The results indicate that the students were more likely to feel confident about their answers in the second test. However, when it comes to metacognition, as assessed by Metacognitive Awareness Inventory (MAI), there was no significant difference between the pretest and posttest (p=.30). In terms of confidence accuracy on the reasoning items, the scores of correctly answered scientific reasoning tests with high confidence (RHC) had increased significantly (M_{pre}= .31, M_{post}=.34, p<.01). This implies that in the second test, a larger number of students were able to answer the reasoning test correctly and with high confidence compared to the first test. On the other hand, students who answered incorrectly with low confidence declined significantly (WLC, M_{pre}= .09, M_{post}=.06, p<.01) in the posttest. Additionally, there was no significant change in incorrect answers with high confidence level (WHC, p=.89).

The various trends of students' answers from the pretest to the posttest are shown in Table 3. Based on Table 3, there were 1365 correct answers and 2735 erroneous answers in the pretest; most of the correct answers had high confidence ratings. The most frequently occurring change was WHC to WHC response (43.85%), showing that there were still incorrect responses with a high degree of confidence in both the pretest and posttest. The second trend was RHC to RHC response (21.61%), which indicated that the students did not change their initial answers that were correctly responded with high confidence. The third trend was WHC to RHC response (10.56%). The RHC to WHC response (8.34%) was ranked fourth, indicating that there were more students who changed their responses from incorrect to correct compared to those who changed their responses from correct to incorrect. The results show that the benefit of changing answers can be seen from the overall outcome. The overall score for the reasoning test in the second test (posttest) was significantly higher than in the pretest. The trend which occurred most was the incorrect change with high confidence either in the pretest or the posttest. It was also noted that the trend of changing from incorrect to correct answer was higher than the change from correct to incorrect.

Table 3

Map of Metacognitive Judgement Change from Pretest to Posttest

Pretest	Posttest	Frequency	Percent (%)
Right Answer, High Confidence (RHC)	RHC	886	21.61
	WHC	342	8.34
	RLC	31	0.76
	WLC	21	0.51
Wrong Answer, Low Confidence (WLC)	WHC	199	4.85
	WLC	96	2.34
	RHC	55	1.34
	RLC	14	0.34
Right Answer, Low Confidence (RLC)	RHC	34	0.83
	WHC	26	0.63
	WLC	13	0.32
	RLC	12	0.29
Wrong Answer, High Confidence (WHC)	WHC	1798	43.85
	RHC	433	10.56
	WLC	112	2.73
	RLC	28	0.68

Answer Modification and Its Relation to Item Difficulty

Table 4 provides insights into the relationship between item difficulty and the trend in answer change. The item difficulty data was obtained from (3PL) IRT model analysis using R studio. In the case of easy questions, students were more likely to change their answers from incorrect to correct rather than from correct to incorrect, although there is also a trend of changing from correct to incorrect answers for item 17 compared to the other easy questions. The most frequently occurring trend for easy questions, as indicated by the data, was retaining the initial response of correct answer with high confidence (RHC to RHC).

Table 4

Overview of Answer Changing Trend in Each Item

Item	Item Difficulty	Most Occuring Trend	Most Occuring Trend (%)	Answer Change (%)	Incorrect to Correct Change (%)	Correct to Incorrect Change (%)	Correct to Correct Change (%)	Incorrect to Incorrect Change (%)
r13	3.13	WHC to WHC	67.32	25.37	10.73	4.39	6.83	78.05
r14	2.46	WHC to WHC	40.00	60.00	9.27	11.71	8.29	70.73
r9	2.33	WHC to WHC	65.85	48.78	8.29	8.78	5.37	77.56
r12	1.94	WHC to WHC	44.39	46.34	9.27	10.24	7.80	72.68
r15	1.83	WHC to WHC	59.51	33.66	13.17	6.83	13.17	66.83
r2	1.55	WHC to WHC	56.59	60.98	15.12	15.61	7.32	61.95
r1	1.55	WHC to WHC	60.49	26.34	11.22	8.78	19.02	60.98
r10	1.53	WHC to WHC	55.61	45.37	13.66	13.66	9.76	62.93
r19	1.45	WHC to WHC	44.39	27.32	14.15	2.44	24.39	59.02
r16	1.28	WHC to WHC	40.00	53.17	14.63	10.24	17.07	58.05
r18	1.26	WHC to WHC	54.15	44.88	11.22	13.17	14.63	60.98
r8	0.98	WHC to WHC	30.73	54.63	16.10	14.63	18.05	51.22
r4	0.79	WHC to WHC	50.73	42.44	10.73	14.15	20.98	54.15
r3	0.66	WHC to WHC	31.22	45.85	18.54	9.27	30.73	41.46
r5	0.46	WHC to WHC	36.59	43.90	16.10	11.22	31.22	41.46
r6	0.30	WHC to WHC	37.07	29.76	12.68	10.24	36.10	40.98
r11	0.17	RHC to RHC	43.90	21.95	11.71	3.41	45.85	39.02
r20	-0.02	RHC to RHC	40.00	35.61	17.07	9.27	43.41	30.24
r7	-0.37	RHC to RHC	52.20	24.88	13.66	5.85	53.66	26.83
r17	-0.84	RHC to RHC	51.22	34.63	11.22	12.20	56.10	20.49

Furthermore, identifying the relationship between item difficulty and answer change is shown in Table 5. As indicated in Table 5, item difficulty shows a meaningful correlation with incorrect-to-incorrect

change, indicating that the more difficult the items, the more likely those items would be answered incorrectly during the pretest and again during the posttest (r=.97, p<.01). It was found that the easier the item, the more likely that students retained the correct answer (r= -.92, p<.01). However, there was no correlation between item difficulty and successful change, from incorrect-to-correct (r=-.42, p=.06) or spoiled answers, from correct-to-incorrect (r=-.07, p=.77).

Table 5

Correlation Table between Item Difficulty and Trend in Answer Changing

	Item Difficulty (1)	Incorrect to Correct Change (2)	Correct to Incorrect Change (3)	Correct to Correct Change (4)	Incorrect to Incorrect Change (5)
(1)	—				
(2)	-0.42	—			
(3)	-0.07	0.08	—		
(4)	-.92**	0.32	-0.26	—	
(5)	.97**	-.49*	0.01	-.96**	—

DISCUSSION

Overconfidence among Indonesian Students in Repeated Scientific Reasoning Test

The objective of this research was to explore Indonesian university students' metacognitive processes, specifically their confidence accuracy in answering scientific reasoning tests, and their answer change between the pretest and posttest. Further, we also explored the relationship between answer changing on scientific reasoning tests by determining test item difficulty. The calibration of confidence accuracy and correctness or incorrectness of reasoning score into four characters (i.e RHC-right answer with high confidence, WHC-wrong answer with high confidence, RLC-right answer with low confidence, WLC-wrong answer with low confidence) was first examined and then the most frequently occurring change from the pretest to posttest was measured.

The results suggest that giving students a chance to redo their scientific reasoning tests could be advantageous in helping them achieve a better

overall outcome. Apart from the cognitive area, in terms of behavioural traits, it provides students a chance to revise their answers and boosts their self confidence in completing the test. However, there is a need to consider whether the students are being accurately confident about their answers or merely feeling confident about their answers being correct since it is critical to acknowledge that Indonesian students were frequently observed as being overconfident (Rachmatullah & Ha, 2019). In this study, the incorrect but highly confident responses (WHC) account for the most frequently occurring variability of the accuracy categories both in the pretest and the posttest. Moreover, the students were found to be prone to overconfidence in answering the scientific reasoning test even when given a second chance to complete the same test.

Students' Overconfidence in Changing Answers; Judgement that Requires Reflective Thinking

The higher proportion of incorrect to correct answer change discovered in this study is in line with findings by Merry et al. (2021) and Stylianou-Georgiou and Papanastasiou (2017). Students, however, may also benefit by retaining their responses given that the second most frequently occurring trend was that students stuck to their correct answers with high confidence. Thus, the question remains, when is it appropriate to modify a response and when is it appropriate for the student to retain the response? Sometimes, when individuals predict the likelihood of remembering a particular item, they do not directly monitor the strength of the memory trace but instead base their judgements on cues or suggestions, that is, they base it on any variable believed to be associated with learning, knowing, or feeling of uncertainty (Coutinho et al., 2020). Adhering to the initial response may be a wise choice if it is based on reflective thinking rather than belief (Couchman et al., 2016). Therefore, if students choose to modify or adhere to a certain response, the most important thing is that judgement is attributed to deliberate thought, and not based on a widely preconceived idea that adhering to the first response is advantageous.

Prior studies proposed a likely explanation regarding metacognition in answer to the changing situations. Students may have assurance in their confidence judgements if they have adequately studied for a test and have attained a sufficient degree of domain knowledge and understanding (Stylianou-Georgiou & Papanastasiou, 2017).

Regarding confidence rating, reliable confidence judgement is made based on the individual's awareness of their abilities, not on a feeling of rightness (Mata et al., 2013). Whereas, when it comes to a feeling of rightness, an initial intuitive response can impact the judgement process, which can affect the final decision. Good performers may also rely on intuition, but if they can reason beyond their first intuition, this is good confidence (Mata et al., 2013). Overconfidence might occur because of a lack of knowledge (Coutinho et al., 2020b). To objectively assess a good performance, students must first learn what constitutes a good performance. Kruger and Dunning (1999) coined the phrase "dual burden" when it comes to overconfidence among low-performing students. These students lack ability and fail to recognise that they lack capability.

Item Difficulty on Reasoning Test Answer Modification

To answer the third research question, this study explored the answer regarding change correlation to item difficulty. The data indicated that questions with a low item difficulty value or easier items followed the trend from RHC to RHC, while items with a high item difficulty value or more difficult items followed the most frequently occurring trend i.e., from WHC to WHC. No strong correlation was found between item difficulty and correct-to-incorrect or incorrect-to-correct change, which also concurred with previous findings by Stylaniou-Georgiou and Papanastasiou (2017). Past studies had further reviewed students' propensity to solve easy or difficult items. Metcalfe and Kornell (2005) observed that people tend to focus on least well-known items, that is, on more difficult item(s). Furthermore, if students are already familiar with the item, there is a propensity to avoid studying it. One thing that Indonesian teachers or researchers need to consider is that monitoring the accuracy itself is not an easy task for Indonesian students. Even for the hardest questions, students are still relatively highly confident that the answers are correct.

Previous studies, which examined similar approaches, reported that question difficulty may have an impact on answer changing. In the current study, the participants were more likely to have low scores on the reasoning test, and thus they will be less likely to make successful answer modifications to such items (Stylaniou-Georgiou & Papanastasiou, 2017). Not all incorrect answers were successfully changed to correct ones in this study. Although, answer changing did not successfully correct the incorrect answer that came from a

confusing question, researchers suggested changing a few answers when given the opportunity to do so in order to improve the overall score (Higham & Gerrard, 2005) with deliberate and reflective thinking (Couchman et al., 2016). Generally, a bias can occur in first-instinct fallacy, in which students keep their answers because of the fear of making a correct-to-incorrect modification or a spoiled response, and they feel that their first choice is more likely to be accurate and changing them will adversely affect their performance (Stone et al., 2022). Stone et al. (2022) summarised that changing the answer also gives rise to weighing psychological costs, such as regret. This idea may be perpetuated because it is more memorable for students to spoil an answer than it is to correct an answer (Kruger et al., 2005).

Students are more prone to feel they are correct than to be actually correct, which is why it is critical to foster the habit of forming second thoughts (Grant, 2021). Through this research, educational instructors can get a further overview of the reasoning and confidence judgement abilities of Indonesian students. Furthermore, the students' proclivity for responding to the same set of items provides additional evidence that sticking to the initial answer is not always an appropriate basis for decision-making on multiple choice assessments. This study extends to the revision study's findings and emphasises the need to avoid following belief or intuitive thinking, but rather prioritising rational decision-making. Previous studies have described the phenomena of overconfidence bias as a barrier to metacognitive capacity. Overconfidence that is repeatedly observed among Indonesian students indicates a need for educational practice to conduct metacognitive training to promote better self-regulated learning among Indonesian students. Efforts to overcome overconfidence bias, particularly among Indonesian students, have been raised by several researchers, including Rusmana et al. (2020), who highlighted that raising students' awareness of overconfidence bias could help minimise the bias. This study adds to the body of evidence that when it comes to sticking to or changing an answer, it should be based on deliberate thought, and not on a widely held idea that sticking to the initial answer results in a better score.

CONCLUSION

This study emphasizes the need for deliberate thought, rather than relying on intuition when it comes to responding to a test.

Overconfidence was observed among Indonesian students, as the most common variability in accuracy categories which was indicated by incorrect answers accompanied with highly confident responses (WHC). Even when given a second chance, the students tended to overestimate their confidence in answering scientific reasoning tests. Regarding answer changes, the study found a higher proportion of incorrect-to-correct changes compared to correct-to-incorrect changes, which is consistent with previous researches. However, students could also benefit from sticking to their initial correct answers. Thoughtful thinking should guide their decision whether to modify or retain a response, rather than blindly adhering to the idea that the initial answer is always advantageous.

ACKNOWLEDGEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

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

Archer, N. S., & Pippert, R. (1962). Don't change the answer! An expose of the perennial myth that first choices are always the correct ones. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 37*, 39–41. https://doi.org/10.1080/00098655.1962.11476207

Ballance, C. T. (1977). Students' expectations and their answer-changing behavior. *Psychological Reports, 41*(1), 163–166. https://doi.org/10.2466/pr0.1977.41.1.163

Benjamin, L. T., Jr., Cavell, T. A., & Shallenberger, W. R. (1984). Staying with initial answers on objective tests: Is it a myth? *Teaching of Psychology, 11*, 133–141. https://doi.org/10.1177/009862838401100303

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *The Journal of Experimental Education*, *69*(2), 133–151. https://doi.org/10.1080/00220970109600653

Brown, G., Irving, E., & Keegan, P. (2008). An introduction to educational assessment, measurement and evaluation. Auckland, NZ: Pearson Education.

Cohen, L., Manion, L., & Morrison, K. (2002). *Research methods in education*. Routledge.

Couchman, J. J., Miller, N. E., Zmuda, S. J., Feather, K., & Schwartzmeyer, T. (2016). The instinct fallacy: The metacognition of answering and revising during college exams. *Metacognition & Learning*, *11*, 171–185. https://doi.org/10.1007/s11409-015-9140-8

Coutinho, M. V., Papanastasiou, E., Agni, S., Vasko, J. M., & Couchman, J. J. (2020). Metacognitive monitoring in test-taking situations: A cross-cultural comparison of college students. *International Journal of Instruction*, *13*(1), 407. https://doi.org/10.29333/iji.2020.13127a

Coutinho, M. V., Thomas, J., Lowman, I. F., & Bondaruk, M. V. (2020b). The Dunning-Kruger effect in Emirati college students: Evidence for generalizability across cultures. *International Journal of Psychology and Psychological Therapy*, *20*(1), 29–36.

Coutinho, M. V., Thomas, J., Alsuwaidi, A. S., & Couchman, J. J. (2021). Dunning-Kruger effect: Intuitive errors predict overconfidence on the cognitive reflection test. *Frontiers in Psychology*, *12*, 1040. https://doi.org/10.3389/fpsyg.2021.603225

Craig, K., Hale, D., Grainger, C., & Stewart, M. E. (2020). Evaluating metacognitive self-reports: Systematic reviews of the value of self-report in metacognitive research. *Metacognition and Learning*, *15*(2), 155–213. https://doi.org/10.1007/s11409-020-09222-y

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Friedman-Erickson, S. (1994). *To change or not to change: The multiple-choice dilemma*. Proceedings of the Annual Institute of the American Psychological Society on the Teaching of Psychology, Washington, DC.

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. https://doi.org/10.25656/01:7485

Grant, A. (2021). *Think again: The power of knowing what you don't know.* Viking Books.

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. https://doi.org/ 10.33225/jbse/21.20.906

Harrison, G. M., & Vallin, L. M. (2018). Evaluating the metacognitive awareness inventory using empirical factor-structure evidence. *Metacognition and Learning*, *13*, 15–38. https://doi.org/10.1007/s11409-017-9176-z

Higham, P. A., & Gerrard, C. (2005). Not all errors are created equal: Metacognition and changing answers on multiple-choice tests. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *59*(1), 28. https://doi.org/10.1037/h0087457

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

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kruger, J., Wirtz, D., & Miller, D. T. (2005). Counterfactual thinking and the first instinct fallacy. *Journal of Personality and Social Psychology*, *88*(5), 725. https://doi.org/10.1037/0022-3514.88.5.725

Lundeberg, M., & Mohan, L. (2009). Context matters: Gender and cross-cultural differences in confidence. In *Handbook of metacognition in education* (pp. 221–239). Routledge.

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology*, *105*(3), 353. https://doi.org/10.1037/a0033640

May, K., & Jackson, T. S. (2005). IRT item parameters and the reliability and validity of pretest, posttest, and gain scores. *International Journal of Testing*, *5*(1), 63–73.

Merry, J. W., Elenchin, M. K., & Surma, R. N. (2021). Should students change their answers on multiple choice questions? *Advances in Physiology Education*, *45*(1), 182–190. https://doi.org/10.1152/advan.00090.2020

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language*, *52*(4), 463–477. https://doi.org/10.1016/j.jml.2004.12.001

Papanastasiou, E. C. (2015). Psychometric changes on item difficulty due to item review by examinees. *Practical Assessment, Research & Evaluation*, *20*(3), 1–10. https://doi.org/10.7275/jcyv-k456

Pieschl, S. (2009). Metacognitive calibration—an extended conceptualization and potential applications. *Metacognition and Learning*, *4*(1), 3–31. https://doi.org/10.1007/s11409-008-9030-4

Rachmatullah, A., & Ha, M. (2019). Examining high-school students' overconfidence bias in biology exam: A focus on the effects of country and gender. *International Journal of Science Education*, *41*(5), 652–673. https://doi.org/10.1080/09500693.2019.1578002

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

Rutherford, T. (2017). The measurement of calibration in real contexts. *Learning and Instruction*, *47*, 33–42. https://doi.org/10.1016/j.learninstruc.2016.10.006

Schraw, G., Kuch, F., & Gutierrez, A. P. (2013). Measure for measure: Calibrating ten commonly used calibration scores. *Learning and Instruction*, *24*, 48–57. https://doi. org/10.1016/j.learninstruc.2012.08.007

Stone, C., Mattingley, J. B., & Rangelov, D. (2022). On second thoughts: Changes of mind in decision-making. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2022.02.004

Stylianou-Georgiou, A., & Papanastasiou, E. C. (2017). Answer changing in testing situations: The role of metacognition in deciding which answers to review. *Educational Research and*

Evaluation, *23*(3–4), 102–118. https://doi.org/10.1080/13803611.2017.1390479

Wagner, D., Cook, G., & Friedman, S. (1998). Staying with their first impulse? The relationship between impulsivity/reflectivity, field dependence/field independence and answer changes on a multiple-choice exam in a fifth-grade sample. *Journal of Research and Development in Education*, *31*, 166–175.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.