Psychological Topics, 28 (2019), 1, 93-113
Original Scientific Paper UDC: 159.955.072
doi:https://doi.org/10.31820/pt.28.1.5

Performance and Metacognition in Scientific Reasoning: The Covariation Detection Task

Pavle Valerjev
University of Zadar, Department of Psychology, Zadar, Croatia

Marin Dujmović
University of Bristol, School of Psychological Science, Bristol, United Kingdom

Abstract

The aim of this study was to introduce a modified version of the covariation detection task to the meta-reasoning framework. This task has been used to assess scientific reasoning through the evaluation of fictitious experiment outcomes and hypothesis testing. The traditional covariation detection task was modified to include only the magnitude versus ratio-bias. The participants' task was to evaluate the effectiveness of an experimental manipulation in a series of fictitious experiments. Experiment 1 (N = 61) consisted of twenty covariation detection tasks. In half of the tasks, normative and heuristic responses were congruent, and for the other half they were incongruent. Experiment 2 (N = 48) had the same experimental design, however, the fictitious data was modified to increase the relative strength of the normative response. After each trial participants provided a judgment of confidence. Results confirmed that the main manipulation of congruence was successful. Participants were more accurate, faster and more confident in the congruent condition. The manipulation from Experiment 2 had a larger impact on response times than on confidence judgments and accuracy. Correct responses were faster in Experiment 2 when compared to Experiment 1, with higher confidence for correct congruent responses. Analyses by response type revealed large individual differences in the relative strength of the processes which generate normative and biased responses. Participants were faster and more confident when rationalizing in favour of their dominant response while they were slower and less confident when decoupling from that dominant response. The covariation detection task provides new valuable insight into meta-reasoning processes.

Keywords: meta-reasoning, scientific reasoning, covariation detection task, cognitive decoupling, cognitive bias, dual-process theory

✉ Pavle Valerjev, Department of Psychology, University of Zadar, Obala kralja Petra Krešimira IV 2, 23000 Zadar, Croatia. E-mail: *valerjev@unizd.hr*

Introduction

The dual-process approach to thinking assumes that there are two distinct categories of processes which lead human reasoning. Traditionally, Type 1 processes have been described as fast, intuitive, based on heuristics and with low cognitive resource requirements. On the other hand, Type 2 processes have been described as slow, analytic, deliberate and cognitively costly. Decades of research have revealed that only the difference in cognitive load remains as a defining feature, while the others are more appropriately viewed as correlates (for an excellent review see Evans & Stanovich, 2013). Modern dual-process models assume that multiple Type 1 processes generate responses during reasoning. These processes may, or may not, generate the same response. A congruent situation is observable when multiple processes generate the same response and an incongruent situation when they generate different responses. The strength of the processes, and thus, the generated responses are not necessarily equal which leads to a relative dominance of one process over the other. If the generated responses are different, then there is the possibility that a conflict between them may be detected. The detection of conflict is presumed to be one of the major triggers for initiation of Type 2 processes. The resolution of this conflict represents Type 2 processing. The conflict may be resolved in one of two ways. First, the dominant response may be chosen as the final response regardless of the conflict. This process is defined as rationalization (Pennycook, Fugelsang, & Koehler, 2015; Wason & Evans, 1975). Second, the alternative response may be chosen as the final response after cognitive dis-attachment from the dominant initial response. This process is defined as cognitive decoupling (Pennycook et al., 2015; Stanovich, 2009a). Additionally, if none of the initial responses are deemed appropriate, further processing is required to generate more responses (processing traditionally labelled as analytical thinking). For a detailed review of the current state of the dual-process approach, see De Neys (2018).

During the past couple of decades, there has been an increasing interest in expanding and supplementing reasoning research by investigating the accompanying metacognitive processes. For this purpose, the framework of meta-reasoning has been developing and growing in significance within the reasoning community. The goal of meta-reasoning research is to provide insight into control and monitoring metacognitive processes and their relationship with reasoning (Ackerman & Thompson, 2015, 2017). Meta-reasoning has been developed following the tradition of meta-memory research (Nelson & Narens, 1990), and adopts some of the measures from that tradition. A number of metacognitive indicators may be measured before (e.g. judgment of solvability), during (e.g. feeling of rightness) or after reasoning processes have been completed (e.g. final judgment of confidence). One of the main findings from meta-reasoning research has shown that participants rarely have insight into normative accuracy when generating confidence judgments (Bajšanski, Močibob, & Valerjev, 2014; Dujmović & Valerjev, 2018; Thompson & Johnson, 2014; Thompson, Prowse Turner, & Pennycook, 2011; Valerjev &

Dujmović, 2017). More salient indicators have been shown to better predict metacognitive judgments. Response times, which provide a measure of response fluency, are a robust predictor of metacognitive judgments (Ackerman & Zalmanov, 2012; Thompson, Evans, & Campbell, 2013; Thompson, Prowse Turner et al., 2013). Our recent work has shown that conflict detection and resolution have an impact on the formation of metacognitive judgments, such as the confidence judgment, independently of fluency (Dujmović & Valerjev, 2018). Furthermore, lower confidence and feeling of rightness are correlated with the probability of rethinking initial answers in reasoning tasks (Shynkaruk & Thompson, 2006; Thompson & Johnson, 2014; Thompson, Prowse Turner et al., 2013). Finally, more emerging evidence points towards large individual differences in sensitivity to conflict (Dujmović & Valerjev, 2018; Frey, Johnson, & De Neys, 2018; Mevel et al., 2015). These findings have proven valuable for developing reasoning models, especially within the dual-process approach, and in understanding the relationship between reasoning and metacognitive processes.

Meta-reasoning research has mostly relied on tasks which are based on processes which may produce conflicting responses based on the experimental manipulation. In incongruent (conflict) versions, these tasks, most commonly, have two dominant responses one of which is based on a cognitive heuristic, while the other is traditionally considered analytical. It is worth noting that modern models of dual processing propose that, what was traditionally considered an analytical response, may rather be viewed as a different type of heuristic. For example, researchers speculate on the existence of *logical intuitions* (De Neys, 2012, 2014) which provide logically correct responses as fast as belief bias. Some of the common tasks are categorical syllogisms (Thompson & Johnson, 2014; Thompson & Morsanyi, 2012), conditional reasoning (Markovits, Thompson, & Brisson, 2015), the base rate neglect task (De Neys & Glumicic, 2008; Dujmović & Valerjev, 2018; Pennycook et al., 2015), items from the Cognitive Reflection Task (CRT) (Toplak, West, & Stanovich, 2014), the Linda problem (Aczel, Szollosi, & Bago, 2016), the denominator neglect task (Kirkpatrick & Epstein, 1992; Thompson & Johnson, 2014), moral reasoning tasks (Bialek & De Neys, 2016), and causal reasoning tasks (Fugelsang & Thompson, 2003; Thompson & Johnson, 2014). Many of the results in these studies point towards convergent conclusions, but it is important to continue investigating new tasks within the meta-reasoning approach for a number of reasons. First, the cognitive bias which is conflicted with normatively correct thinking can vary depending on the task. This means that investigating new tasks and biases provides valuable information about the underlying cognitive mechanisms. Additionally, converging evidence connects traditionally separated domains of reasoning such as inductive reasoning, deductive reasoning, probabilistic reasoning, moral reasoning and others. Second, it is difficult to find a fine balance between methodological control, experimental manipulation and ecological validity. For example, the commonly used version of the base rate neglect task (De Neys & Glumicic, 2008) formally does not have a normatively correct response. Formal

logic, on the other hand, has a normatively correct response, but seems artificial to participants. Tasks such as the Linda problem rely on the conjunction rule which is rather difficult to grasp. Third, in most of these tasks the dominant Type 1 response is far more dominant than the alternative (Dujmović & Valerjev, 2018; Pennycook et al., 2015; Thompson et al., 2011; Thompson, Prowse Turner et al., 2013). This means that experimental manipulations have to be implemented in an extreme form in order to overcome this gap and may not provide enough in-depth insight into the process which generates the alternative response. It also makes it more difficult to examine the sensitivity to the conflict between the two responses when it is present.

The reasoning literature has an abundance of tasks which have traditionally been used to investigate human rationality. These tasks cover a variety of different areas of reasoning. Scientific reasoning includes the evaluation of evidence, hypothesis formation and testing, evaluating causal relationships etc. (Stanovich, 2009b). These skills are used in everyday thinking as well as in a number of laboratory experiments. Some of the tasks which employ scientific reasoning are the Wason selection task, the covariation detection task, causal reasoning tasks and various probabilistic reasoning tasks. The covariation detection task has been described as an exemplary task concerned with scientific reasoning (Stanovich, West, & Toplak, 2016). Toplak, West, and Stanovich (2011) investigated the relationship between the CRT and what they labelled as heuristics-and-biases tasks. The covariation detection task which has been adopted for the current study was one of those tasks. The original task was described as shown below.

A doctor had been working on a cure for a mysterious disease. Finally, he created a drug that he thinks will cure people of the disease. Before he can begin to use it regularly, he has to test the drug. He selected 300 people who had the disease and gave them the drug to see what happened. He selected 100 people who had the disease and did not give them the drug in order to see what happened. The table below indicates what the outcome of the experiment was:

	Cure	
	Yes	No
Treatment present	200	100
Treatment absent	75	25

Participants were asked to judge whether this treatment was positively or negatively associated with the cure for this disease by circling a number from a scale ranging from -10 (*strong negative association*) to +10 (*strong positive association*). Negative judgments, which indicated the inefficacy of the treatment, were scored as correct.

This task presents participants with a simple fictitious experiment and asks them to evaluate the results. Early research in covariation tasks was concerned with how people process covariations and how they valued information based on spatial layout

of the data rather than researching biases in reasoning (Kao & Wasserman, 1993; Wasserman, Dorner, & Kao, 1990). Later, Stanovich, and West (1998a, 1998b) introduced a belief bias component to the task. As can be seen in the example above, belief in the effectiveness of a cure compared to a placebo needs to be ignored in order to recognize that the cure was inefficient (Sa, West, & Stanovich, 1999). The ratio of positive to negative outcomes is higher in the treatment-absent condition when compared to the treatment condition. The task requires decontextualization in order to be solved correctly. While studying this problem we noticed a more basic principle which may be the main source of bias in the specific example above. The principle seems to have the same underlying mechanism as shown in the ratio bias (Miller, Turnbull, & McFarland, 1989) and denominator neglect (Kirkpatrick & Epstein, 1992) tasks. Concrete frequencies are more salient and processed more quickly than ratios which leads to a bias towards the higher frequency of positive outcomes in the treatment when compared to the treatment-absent condition. Additionally, the difference between the positive and negative outcomes in the treatment condition is larger when compared to the treatment-absent condition. It is worth noting that a similar tradition of research has focused on the formation of illusory correlations from experience of event frequencies. In this tradition participants implicitly form correlations between events or characteristics which both occur with a high or both with a low frequency even though they do not have all of the information necessary to asses the true relationship between the variables in question (Fiedler, Kutzner, & Vogel, 2013; Mullen & Johnson, 1990).

Overview of the Experiments

We decided to adapt the task by eliminating the belief bias component and manipulating only the magnitude-versus-ratio conflict. Belief bias was eliminated by producing content in which prior knowledge and experience were not in favour of either option. The main experimental manipulation (the conflict of magnitude versus ratios) was achieved by generating specific sets of frequencies and ratios for our tasks (the process is explained in detail in the Appendix A. With such manipulation, we believe three desired outcomes were achieved. First, the task has a normatively correct, and a biased response which may or may not be in conflict depending on the experimental condition. Second, we expected that adjusting the ratios and frequencies provides an opportunity to achieve different levels of conflict. Third, the task achieves higher levels of ecological validity when compared to many other reasoning tasks. Daily life often requires the evaluation of outcomes based on empirical data.

The general goal of this study was to investigate the modified covariation detection task using the rapid response paradigm within the meta-reasoning framework. This task, as far as we know, has not been researched within the meta-reasoning framework using this paradigm. The specific goal of Experiment 1 was to confirm that conflict can be successfully manipulated and to determine the effects of

this conflict on accuracy, response times and confidence judgments. Experiment 2 was designed to induce a different level of conflict when compared to Experiment 1 by increasing the ratio of positive to negative outcomes for the correct responses. Additionally, the second experiment was intended as a partial replication of Experiment 1 on an independent sample. We expected that participants would be more accurate, faster and more confident in the congruent trials and that these differences would increase in Experiment 2 when compared to Experiment 1.

Method

Participants

Participants for both experiments were recruited among undergraduate psychology students unfamiliar with the task. A total of 109 participants volunteered for this study (61 for Experiment 1 and 48 for Experiment 2).

Materials

The materials for this study were carefully constructed using the following procedure.

Twenty imaginary scenarios which describe a change in a number of different fields were generated. These included new products, production processes, treatments, marketing campaigns etc. Alongside the description of what was introduced, additional information was provided on what was measured in order to judge the success of the innovation. For each scenario, a table of data was presented indicating the efficiency of two options (for example, the new and old techniques). An example of a scenario and data matrix can be seen in Figure 1.

A fertility clinic plans to introduce a new technique for the artificial insemination of human egg cells. They wanted to compare the new technique to the present one. The number of successful fertilizations was recorded for comparison.

	Successful fertilization	Unsuccessful fertilization
New technique	94	31
Old technique	228	109

Figure 1. An example of the data covariation scenario and data matrix.

The example is incongruent because the correct option (the new technique) is accompanied by a smaller number of successful fertilizations and the difference

between successful and unsuccessful fertilizations is smaller when compared to the old technique. However, the new technique is more efficient (a better ratio between successful and unsuccessful fertilizations) and is thus the correct option. A congruent version of this example would assign both a high frequency and high ratio (efficiency) to the same option. Frequencies were generated based on an algorithm which can be found in Appendix A, which resulted in a 3 to 1 ratio of positive when compared to negative outcomes for the correct option, and a 2 to 1 ratio for negative outcomes. The final sets of frequencies were randomly assigned to the twenty scenarios. Finally, in half of the scenarios, both for the congruent and incongruent trials, the change was the correct option, and for the other half it was the incorrect one. It is also important to note that the contents of the scenarios were generated in a way to avoid possible belief bias effects due to prior experience. This was done by not using concrete names of products and processes described in the scenarios as well as picking contexts in which there was no reason to believe the introduced novelty would necessarily lead to an improvement.

For Experiment 2, the only change was the increase of higher ratios in order to achieve an average of 8 to 1. This was achieved by adjusting the negative outcome frequencies for the correct options in Experiment 1. For example, the number of unsuccessful fertilizations in the previous example was lowered to 12.

Procedure

Both experiments were conducted using E-Prime v2.10.356. Each trial started with a fixation cross which lasted 1000 milliseconds. The fixation cross was followed by the scenario. Participants were told to press a key on the keyboard after they carefully read the scenario. After a key was pressed a table similar to the one in Figure 1 appeared. Participants were instructed to analyse the data provided in the table and decide, as fast as possible, whether the change was better than the old/control option. Responses were made by pressing the appropriate key on the keyboard (the "s" key was labelled as NO, and the "k" key as YES). Responses and response times were recorded. After each response, the participants were asked to provide a confidence judgment on a six point scale, ranging from 50% (guessing) to 100% (complete confidence), with 10% increments.

Participants completed a single practice trial before the main measurement. The order of the main trials was randomized for each participant. Further, the position of positive and negative outcomes in the data tables was rotated to avoid bias towards specific positions in the tables. The procedure was the same in both experiments.

Results

Even though the study describes two distinct experiments conducted at different times on independent samples, the data will be analysed in a single section for brevity and in order to make it easier to compare differences between the two experiments.

Total Data Set Analysis

Mean response times for congruent and incongruent conditions were calculated from median values from each participant. Mean confidence judgments were calculated from means for each participant. Finally, mean accuracy was calculated from the percentage of correct responses of each participant. The means and standard deviations from the total data set can be seen in Table 1.

Table 1

Mean (Standard Deviations) of Response Times, Confidence Judgments and Accuracy as a Function of Congruence in Experiments 1 and 2

	Response times [ms]	Confidence [%]	Accuracy [%]
Experiment 1			
Congruent	9703.88 (4246.47)	83.21 (8.18)	92.13 (12.26)
Incongruent	11646.03 (5276.12)	78.07 (10.03)	57.38 (30.16)
Experiment 2			
Congruent	7292.94 (2785.43)	88.32 (7.87)	96.78 (7.21)
Incongruent	10037.78 (3647.91)	79.58 (7.77)	62.92 (31.75)

In order to determine the effects of congruence and ratio extremity (experiment), 2(experiment) × 2(congruence) mixed ANOVAs were conducted on response times and confidence judgments. For accuracy, nonparametric tests were used since the distributions of accuracy data deviated significantly from normal. Results of the ANOVAs can be seen in Table 2.

Table 2

ANOVA Results for Response Times and Confidence Judgments

	Response times		Confidence judgments	
	$F(1, 107)$	η_p^2	$F(1, 107)$	η_p^2
Experiment	6.97^{**}	.06	4.50^{*}	.04
Congruence	76.44^{**}	.42	157.64^{**}	.60
Interaction	2.24	.02	10.55^{**}	.09

$^{*}p < .05$; $^{**}p < .01$.

For both response times and confidence, we found a large effect of congruence. Participants were faster and more confident for responses in congruent when

compared to incongruent trials. A smaller main effect of experiment (ratio extremity) was also significant. Participants were slightly faster and more confident in Experiment 2 when compared to Experiment 1. However, for confidence judgments an interaction effect was significant due to the fact the only significant difference between the experiments was for congruent, but not for incongruent trials.

In order to determine the difference in accuracy depending on congruence, a Wilcoxon matched pairs test was conducted. The result showed participants were significantly more accurate in congruent when compared to the incongruent trials (T = 85.00, Z = 8.05, $p < .01$). Man-Whitney U tests were conducted to determine possible differences between the two experiments (all $U > 1161$, all $Z < 1.84$, $p > .05$). Results showed there were no significant differences between the two experiments for both congruent and incongruent trials.

Analysis by Response Type

Further analyses were conducted in order to determine the effect of response type on response times and confidence judgments. For both experiments, mean response times and confidence judgments were calculated for correct congruent, correct incongruent and incorrect incongruent responses and analysed by conducting 2×3 mixed ANOVAs. Incorrect responses in the congruent version of the task are not cued by the hypothesized heuristic or by deliberate reasoning about proportions so it can be attributed to either response errors or processes not predicted by the design, so they are excluded from these analyses. Results of these analyses can be seen in Table 3.

Table 3

ANOVA Results for Response Times and Confidence Judgments Depending on the Response Type (Correct Congruent, Correct Incongruent and Incorrect Incongruent)

	Response times			Confidence judgments		
	$F(df_1, df_2)$		η_p^2	$F(df_1, df_2)$		η_p^2
Experiment	3.02	(1, 86)	.03	3.59	(1, 86)	.04
Response type	26.96**	(2, 172)	.24	66.05**	(2, 172)	.43
Interaction	1.41	(2, 172)	.02	3.41*	(2, 172)	.04

*$p < .05$; **$p < .01$.

Only the main effect of response type was significant for response times. Post-hoc analysis (Tukey HSD) showed that correct responses in the congruent trials (M = 8622 ms, SD = 3859) were faster when compared to both correct (M = 11839 ms, SD = 4380) and incorrect (M = 12267 ms, SD = 7777) responses for incongruent trials. The difference between the two types of responses in the incongruent condition was not significant. While the main effect of the experiment was marginal (p = .09)

and the interaction effect was not significant, the pattern of results is interesting as can be seen in Figure 2.

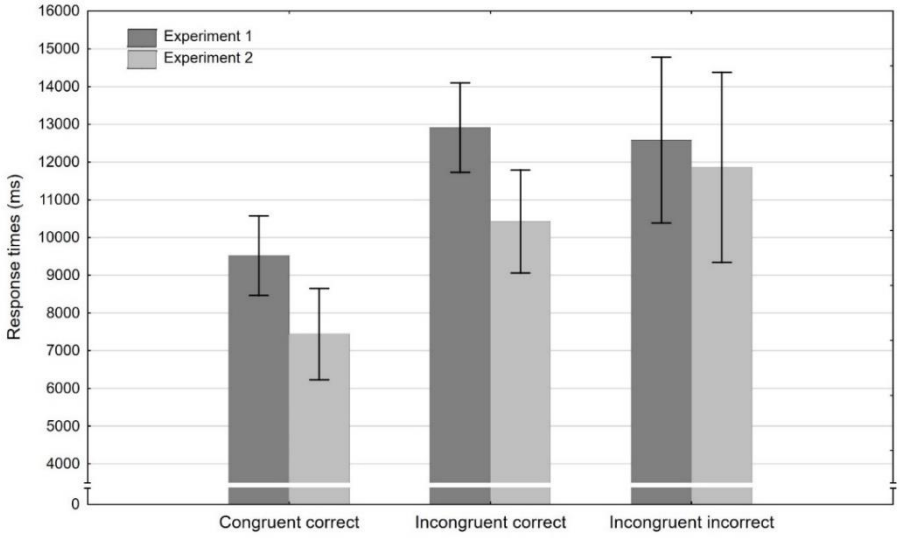

Figure 2. Response times as a function of response type in Experiments 1 and 2 (spreads represent 95% confidence).

For confidence judgments the main effect of response type and the interaction were significant. Post-hoc analysis showed participants were significantly more confident for correct congruent (M = 85.42%, SD = 8.37) than both correct (M = 77.84%, SD = 9.12) and incorrect (M = 75.64%, SD = 10.51) incongruent responses. The difference between the two types of responses in the incongruent condition was also significant. Participants were more confident in Experiment 2 when compared to Experiment 1 only for correct congruent responses, but not for either of the response types in the incongruent condition. The mean values can be seen in Figure 3.

As mentioned in the Method section, it was correct to recognize the change as better when compared to the old/control option for half of the trials, and for half the trials it was correct to recognize it as worse. An additional analysis was conducted by introducing this control feature as an independent variable (new better/old better) which we will label as the *novelty* effect. The same analyses as the ones above were conducted with this additional effect. Two (response times and confidence) 2(experiment) ×2(novelty) × 3(response type) ANOVAs were conducted. Results showed that all of the previously significant effects were the same and did not interact with the novelty effect. Interestingly, for confidence judgments the main effect of novelty was significant ($F(1, 107) = 13.95$, $p < .01$, $\eta_p^2 = .11$). Participants were more confident in trials in which the change was the normatively better choice (M =

82.94%, *SD* = 8.62) than the old/control option (*M* = 81.25%, *SD* = 8.40). The effect was not significant for response times.

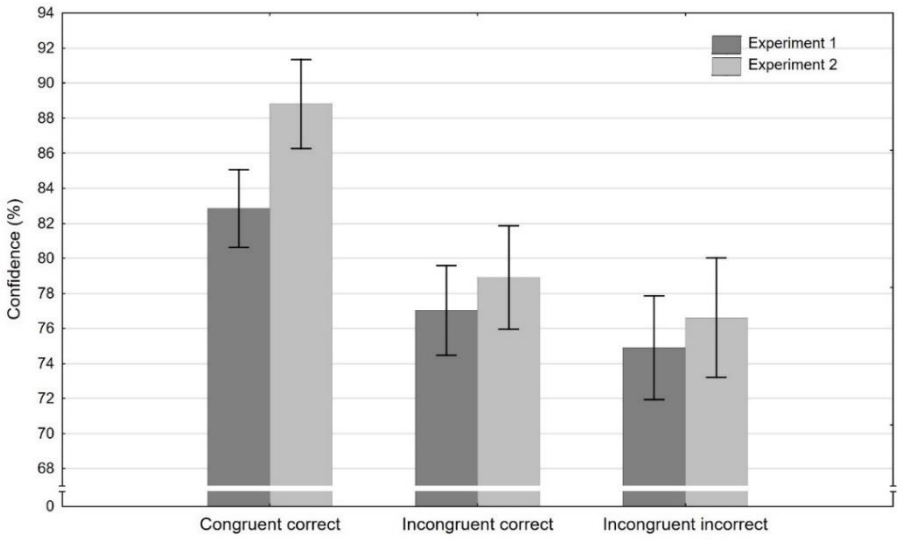

Figure 3. Confidence judgments as a function of response type in Experiments 1 and 2 (spreads represent 95% confidence).

Analysis of Individual Differences

Previous studies have shown there are significant individual differences in conflict sensitivity during reasoning (Dujmović & Valerjev, 2018; Frey et al., 2018; Mevel et al., 2015). Similar to a recent study (Dujmović & Valerjev, 2018) we decided to test for possible differences by forming two sub-groups of participants based on their accuracy in the incongruent condition. Mean accuracy in that condition across both experiments was 59.84% (with a median of 60%) so the median was chosen as the critical value. Participants with a lower accuracy were assigned to the low accuracy group (*N* = 49), while participants with 60% or higher accuracy were assigned to the high accuracy group (*N* = 60). Two 2(experiment) × 2(accuracy group) × 3(response type) ANOVAs were conducted for response times and confidence judgments. Response times analysis revealed (alongside previously determined effects) a main effect of accuracy group ($F(1, 84) = 5.51$, $p < .05$, $\eta_p^2 = .06$) which showed participants in the higher accuracy group were generally slower. However, the key finding was a significant accuracy group by response type interaction ($F(2, 168) = 14.69$, $p < .01$, $\eta_p^2 = .15$) which can be seen in Figure 4.

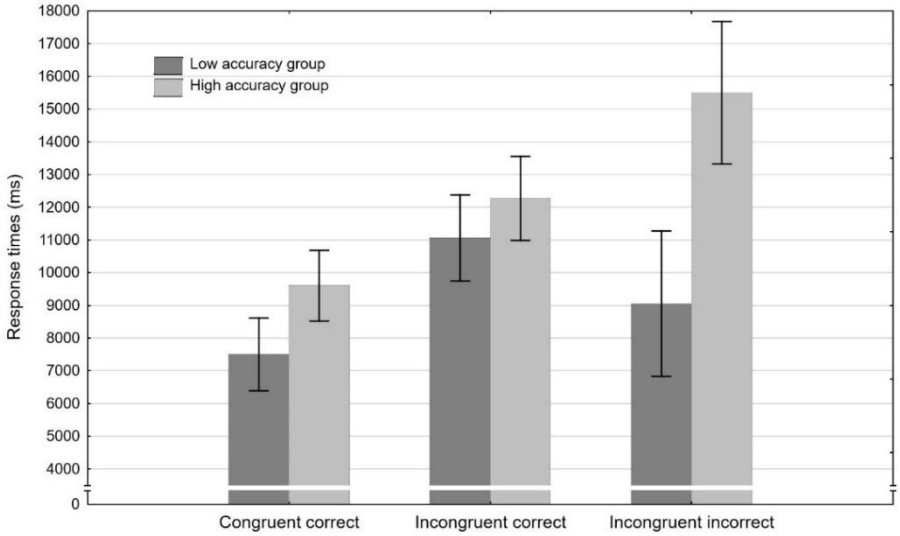

Figure 4. Response times as a function of response type and accuracy group (spreads represent 95% confidence).

Post-hoc analysis of the interaction showed the high accuracy group was significantly slower when compared to the low accuracy group only when responding incorrectly in the incongruent condition, but differences in the remaining two response types were not significant. It is worth noting that the accuracy group by response type effect was the same for both experiments, and that no other interactions were significant.

The same analysis of confidence judgments revealed a significant accuracy group by response type interaction ($F(2, 168) = 12.11$, $p < .01$, $\eta_p^2 = .13$) which can be seen in Figure 5. Post-hoc analysis showed participants in the high accuracy group, in comparison with the low accuracy group, were significantly more confident when providing correct responses in the incongruent condition. The differences were not significant for the other two response types, but it is worth noting that the direction of the difference between the two groups was reversed for incorrect responses in the incongruent condition. As was the case for response times, the accuracy group effect interacted only with response type and the pattern of results was the same in both experiments.

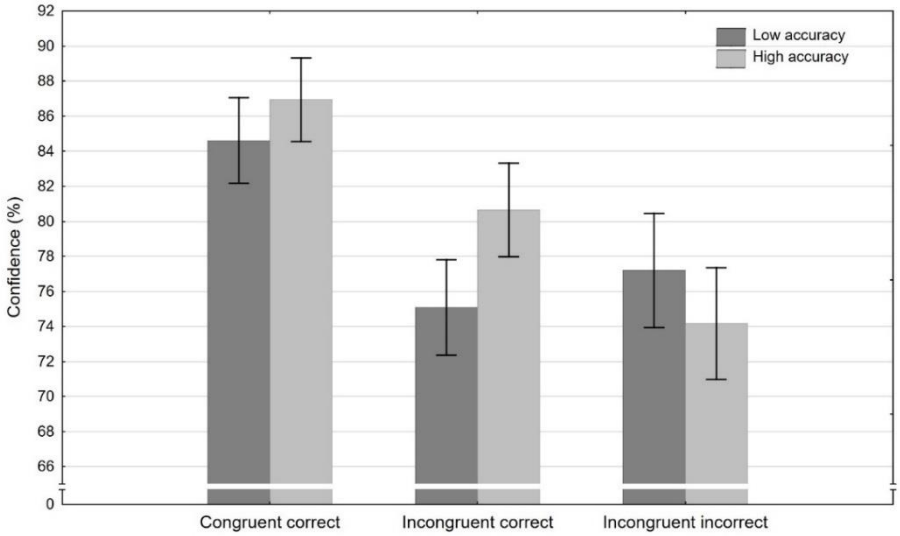

Figure 5. Confidence judgments as a function of response type and accuracy group (spreads represent 95% confidence).

Discussion

The goal of this study was to modify the covariation detection task in order to investigate the magnitude-versus-ratio bias within the meta-reasoning framework, which is the first study of this kind, as far as we know. This task has been used to provide insight into what is described as scientific reasoning which involves testing hypotheses, evaluating results, covariation and causation (Stanovich, 2009b; Stanovich et al., 2016). In general, the two independent experiments revealed a strong effect of conflict on all three of the dependent variables. Participants were more accurate, faster and more confident when the frequency and ratio information were congruent (point towards the same response). Experiment 2 successfully replicated the effect of congruence and revealed a significant effect of ratio extremity on response times and confidence judgments. Participants from Experiment 2 were, generally speaking, faster and more confident than participants from Experiment 1. We also expected ratio extremity to increase accuracy in Experiment 2 when compared to Experiment 1. The data shows an increase of accuracy in both congruent and incongruent conditions in Experiment 2, however, the difference was not statistically significant. From the general analysis, we can conclude that the main experimental manipulation of congruence was successful and had a very similar effect as in other reasoning tasks. When conflict is successfully detected in incongruent trials, additional processing is required to resolve the conflict. This additional processing prolongs response times. Conflict also leads to lower levels of confidence which other studies have shown to be both a direct and mediated effect.

Some results show that the detection of conflict lowers confidence levels even when there is no significant impact on response times (Dujmović & Valerjev, 2018), on the other hand, there is an indirect effect of conflict on confidence due to the well documented relationship between fluency and metacognitive judgments (Ackerman & Zalmanov, 2012; Thompson, Evans et al., 2013; Thompson, Prowse Turner et al., 2013). Conflict resolution prolongs response times which are negatively correlated with confidence judgments. Due to the low proportions of correct responses in previous studies with the covariation detection task (Toplak et al., 2011; West, Toplak, & Stanovich, 2008; Stanovich & West, 1998b), we speculated that the bias towards larger numbers would be the dominant Type 1 process. However, in previous studies, the task was significantly different from our modification in two important ways. First, belief bias was present, and in some cases, represented the main source of bias (Stanovich & West, 1998b; Sa et al., 1999; West et al., 2008) without known interactions with the bias towards larger numbers. Second, participants were instructed to respond in terms of a perceived positive or negative relationship between the treatment and the positive outcomes. Our study showed that, when controlling for other effects, accuracy in incongruent conditions proves to be quite high (around 60%). This would indicate that the two processes which generate the ratio and frequency based responses are closer in strength than expected. It is quite plausible that in this format, the ratio information generated the more dominant response. If this were the case, we would expect the proportion of successful conflict detections in Experiment 2 to be lower than in Experiment 1, leading to more accurate, faster and more confident responses in incongruent trials. While this was generally true, the non-significant difference in confidence represents a deviation from those expectations.

Since simple effects do not provide enough detailed information in these types of studies, it is important to analyse data based on response type. Correct responses in the congruent condition are used as a baseline. These responses should result in highest accuracy rates, shortest response times and highest confidence levels since both the heuristic and normative processing cue the same option. In order to truly test for the effect of conflict and its detection on meta-reasoning, it is important to compare the correct congruent responses to both correct and biased responses in the incongruent condition. For example, if there were no difference between the correct congruent and biased incongruent responses, then it would be difficult to argue that the participants are sensitive to conflict regardless of the response they make in the incongruent condition. Studies within the reasoning and meta-reasoning framework have shown that people are indeed sensitive to conflict even in the presence of strong biases (De Neys & Glumicic, 2008; Dujmović & Valerjev, 2018; Pennycook et al., 2015; Thompson & Johnson, 2014). If the two processes which generated the responses in our version of the task were closer in strength than is usually the case in reasoning research, we would expect a clear effect of conflict regardless of response type in the incongruent condition. The results proved this expectation to be correct. Participants were slower and less confident for both correct and incorrect responses

in the incongruent condition when compared to the congruent condition, and this was true in both experiments. Additionally, participants were more confident when giving correct responses in incongruent trials while the difference in response times was not significant. These results provide an additional argument towards the speculated dominance of ratio versus frequency information in this format of the task. If the ratio information is relatively stronger, then cognitive decoupling of that response, and towards the frequency-based response, should have a greater impact than vice versa. This would explain the observed differences in confidence levels between the correct and incorrect responses in the incongruent condition. The significant difference in confidence and non-significant difference in response times also provide evidence towards an independent effect of conflict detection and resolution on metacognitive judgments, beyond the fluency-metacognition relationship.

Closer observation of the data revealed that participants in Experiment 2 were faster when giving correct responses in congruent and incongruent trials when compared to participants in Experiment 1, but not when giving incorrect responses in incongruent trials. The more interesting analysis of confidence judgments revealed an interaction effect between ratio extremity and response type. Confidence was significantly higher in Experiment 2 for correct congruent responses but not for either correct or incorrect responses in the incongruent condition. We expected that in Experiment 2 confidence would be higher for correct, but lower for incorrect responses in the incongruent condition when compared to Experiment 1. This was expected because the information in favour of the correct response was stronger in Experiment 2, and conversely, the arguments for incorrect responses were weaker, relatively speaking. However, the only significant difference in confidence between the two experiments was observed for correct congruent responses, in favour of Experiment 2. This would indicate that the drop in confidence for incongruent trials is mainly determined by the fact that conflict was detected and by the outcome of the resolution of that conflict (response type). The increased strength of the ratio information in Experiment 2 did not lead to a more extreme drop in confidence for incorrect responses or to a mitigated drop for correct responses in the incongruent condition. Considering sample size and type of measurement, there is a possibility that the manipulation was not extreme enough to reveal the effect. Another possibility is that the increased ratio extremity influences the probability of conflict detection but has a limited impact after successful detection.

Since the complicated task of drawing conclusions about processes within meta-reasoning is made even more complicated by findings of significant individual differences in both conflict sensitivity (Dujmović & Valerjev, 2018; Frey et al., 2018; Mevel et al., 2015) and cognitive styles (Stanovich et al., 2016; Stanovich & West, 1998a; West et al., 2008), it is necessary to incorporate them into meta-reasoning experiments even when those differences are not the focus of the study. In a previous study (Dujmović & Valerjev, 2018), we demonstrated a successful method of

differentiating between conflict sensitive and conflict insensitive participants based on their accuracy in incongruent trials. For this study, participants were grouped based on median accuracy in the incongruent trials (C = 60%) into a low and high accuracy group. Response time and confidence data analysis based on response type and accuracy group revealed, perhaps, the most interesting findings of the study. Results showed participants in the high accuracy group were slower and less confident for incorrect responses when compared to correct ones in the incongruent condition. On the other hand, participants in the low accuracy group were faster and more confident for incorrect when compared to correct responses in the incongruent condition. This finding indicates that the dominance pattern of the processes which generate the two types of responses is not the same for these two groups of participants. In the high accuracy group, the ratio carried more weight relative to the magnitude information. Conversely, in the low accuracy group, the pattern was reversed. The result has two very important implications in light of recent developments within the dual-process approach to reasoning as well as for the relationship between reasoning and metacognitive processes. First, previous studies have shown a trend which would indicate that cognitive decoupling is slower and leads to lower confidence relative to rationalization (Dujmović & Valerjev, 2018; Pennycook et al., 2015). Our results reveal a significant effect, but this was achieved only because the participants were differentiated into two accuracy groups. For both groups, the decoupling process prolonged response times and lowered confidence levels when compared to rationalization. Second, responses which would traditionally be attributed to purely analytical thinking may, for some participants, reflect processing similar to heuristic reasoning based on experience, automatization and other factors. This provides strong validation for the new hybrid models of dual-processing which propose the existence of multiple Type 1 processes, including processing traditionally attributed to Type 2 systems.

The interesting result labelled as the novelty effect provides material for the final discussion topic of this study. Results showed a slight systematic increase of confidence for trials in which the novel treatment/product/process was normatively the correct response when compared to trials in which the old treatment/ product/process was the correct response. The current study does not provide enough statistical power to detect subtle interactions which include the novelty effect but may prove interesting for a different study. We assume that the effect may be a specific type of belief bias. It may be the case that people expect changes and introductions of new treatments/products/processes to be more efficient when compared to current ones. This would easily be explained by the very nature of technological and societal development. Even when a product fulfils its purpose to a satisfactory degree, changes commonly lead to optimisation and new functions. This everyday experience may have led to the formation of such a specific, though slight, form of belief bias.

While the study mostly confirmed expectations, the noticeable absence of certain effects and interactions when trends do exist can be attributed to methodological issues. First, there was no pre-study to determine whether the scenarios were balanced and completely devoid of prior belief bias which could have interacted with the main experimental manipulations on some but not all trials. Second, there were only ten trials per experimental condition which in turn probably lead to lower power to detect differences between the two experiments, and especially when analysing differences depending on response type. Additional experimental control and an expanded set, as well as a larger sample, would be recommended for further use of the task with the same paradigm as developed in this study.

To summarize, the current study provided strong evidence for the effect of conflict on response times and metacognitive judgments in a modified covariation detection task which tests aspects of scientific reasoning. The main experimental manipulation exploited a known bias towards magnitudes in probabilistic reasoning to induce conflict, which in turn lowered accuracy, prolonged responses and lowered confidence when compared to a congruent version of the task. Furthermore, cross-experimental analyses revealed that increasing the ratio extremity led to faster responses for correct congruent and correct incongruent responses, and higher confidence only for correct congruent responses. Finally, findings show that there are large individual differences in the relative strength of the Type 1 processes which generated responses in the task. When these differences are included into the statistical analysis, a strong pattern of results show that cognitive decoupling has a greater impact on both response times and metacognitive judgments relative to rationalization. The modified covariation detection task may prove to be a valuable tool for meta-reasoning research in the future since it allows manipulations of multiple factors which influence reasoning processes. Scientific reasoning represents everyday thinking and reasoning to a higher degree, compared to other, more artificial, aspects and tasks within reasoning research. For this reason, it should be incorporated and studied in greater detail from a meta-reasoning perspective.

References

Ackerman, R., & Thompson, V. A. (2015). Meta-reasoning: What can we learn from meta-memory? In A. Feeney & V. A. Thompson (Eds.), *Reasoning as memory* (pp. 164-182). New York: Psychology Press.

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

Ackerman, R., & Zalmanov, H. (2012). The persistence of the fluency-confidence association in problem solving. *Psychonomic Bulletin & Review, 19*, 1187-1192. https://doi.org/10.3758/s13423-012-0305-z

Aczel, B., Szollosi, A., & Bago, B. (2016). Lax monitoring versus logical intuition: The determinants of confidence in conjunction fallacy. *Thinking & Reasoning, 22*(1), 99-117. https://doi.org/10.1080/13546783.2015.1062801

Bajšanski, I., Močibob, M., & Valerjev, P. (2014). Metacognitive judgments and syllogistic reasoning. *Psychological Topics, 23*(1), 143-165.

Bialek, M., & De Neys, W. (2016). Conflict detection during moral decision-making: Evidence for deontic reasoners' utilitarian sensitivity. *Journal of Cognitive Psychology, 28*(5), 631-639. https://doi.org/10.1080/20445911.2016.1156118

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science, 7*(1), 28-38. https://doi.org/10.1177%2F1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning, 20*(2), 169-187. https://doi.org/10.1080/13546783.2013.854725

De Neys, W. (Ed.) (2018). *Dual process theory 2.0*. New York: Routledge.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*, 1248-1299. https://doi.org/10.1016/j.cognition.2007.06.002

Dujmović, M., & Valerjev, P. (2018). The influence of conflict monitoring on meta-reasoning and response times in a base rate task. *Quarterly Journal of Experimental Psychology, 71*(12), 2548-2561. https://doi.org/10.1177%2F1747021817746924

Evans, J. St. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223-241. https://doi.org/10.1177/1745691612460685

Fiedler, K., Kutzner, F., & Vogel, T. (2013). Pseudocontingencies: Logically unwarranted but smart inferences. *Current Directions in Psychological Science, 22*(4), 324-329. https://doi.org/10.1177/0963721413480171

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology, 71*(5), 1188-1208. https://doi.org/10.1080/17470218.2017.1313283

Fugelsang, J. A., & Thompson, V. A. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition, 31*, 800-815. https://doi.org/10.3758/BF03196118

Kao, S. F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(6), 1363-1386. http://doi.org/10.1037/0278-7393.19.6.1363

Kirkpatrick, L. A., & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Further evidence for two conceptual systems. *Journal of Personality and Social Psychology, 63*(4), 534-544. https://doi.org/10.1037/0022-3514.63.4.534

Markovits, H., Thompson, V. A., & Brisson, J. (2015). Metacognition and abstract reasoning. *Memory & Cognition, 43*, 681-693. https://doi.org/10.3758/s13421-014-0488-9

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology, 27*(2), 227-237. https://doi.org/10.1080/20445911.2014.986487

Miller, D. T., Turnbull, W., & McFarland, C. (1989). When a coincidence is suspicious: The role of mental simulation. *Journal of Personality and Social Psychology, 57*(4), 581-589. https://doi.org/10.1037/0022-3514.57.4.581

Mullen, B., & Johnson, C. (1990). Distinctiveness-based illusory correlations and stereotyping: A meta-analytic integration. *British Journal of Social Psychology, 29*, 11-27. https://doi.org/10.1111/j.2044-8309.1990.tb00883.x

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. Brower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 26, pp. 125-173). San Diego: Academic Press.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*, 34-72. https://doi.org/10.1016/j.cogpsych.2015.05.001

Sa, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*(3), 497-510. https://doi.org/10.1037/0022-0663.91.3.497

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition, 34*(3), 619-632. https://doi.org/10.3758/BF03193584

Stanovich, K. E. (2009a). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In K. Frankish & J. St. B. T. Evans (Eds.), *In two minds: Dual processes and beyond* (pp. 55-88). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0003

Stanovich, K. E. (2009b). *What intelligence tests miss: The psychology of rational thought.* New Haven: Yale University Press.

Stanovich, K. E., & West, R. F. (1998a). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*(2), 161-188. http://dx.doi.org/10.1037/0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (1998b). Who uses base rates and P(D/~H)? An analysis of individual differences. *Memory & Cognition, 26*(1), 161-179. https://doi.org/10.3758/BF03211379

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient. Toward a Test of rational thinking*. Cambridge, MA: The MIT Press.

Thompson, V. A., Evans, J. St. B. T., & Campbell, J. I. D. (2013). Matching bias on the selection task: It's fast and feels good. *Thinking & Reasoning, 19*(3), 431-452. https://doi.org/10.1080/13546783.2013.820220

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning, 20*(2), 215-244. https://doi.org/10.1080/13546783.2013.869763

Thompson, V. A., & Morsanyi, K. (2012). Analytic thinking: Do you feel like it? *Mind & Society, 11*, 93-105. https://doi.org/10.1007/s11299-012-0100-6

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*, 107-140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Thompson, V. A., Prowse Turner, J. A., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*, 237-251. https://doi.org/10.1016/j.cognition.2012.09.012

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*, 1275-1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20*(2), 147- 168. https://doi.org/10.1080/13546783.2013.844729

Valerjev, P., & Dujmović, M. (2017). Instruction type and believability influence on metareasoning in a base rate task. In: G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 3429-3434). Austin, TX: Cognitive Science Society.

Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition, 3*, 141-154. http://dx.doi.org/10.1016/0010-0277(74)90017-1

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of interevent contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(3), 509-521. http://doi.org/10.1037/0278-7393.16.3.509

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology, 100*(4), 930-941. http://doi.org/10.1037/a0012842

Received: October 14, 2018

Appendix A

The frequencies for positive and negative outcomes in each scenario were assigned by adhering to the following algorithm:

1. The correct option had to have a 3 to 1 ratio of positive when compared to negative outcomes, while the incorrect option had to have a 2 to 1 ratio of positive to negative outcomes.
2. To ensure the participants would not encounter the exact same ratios and frequencies, twenty random ratios ranging from 2.8-3.3 to 1 (with an average of 3 to 1), and twenty ratios ranging from 1.8-2.1 to 1 (with an average of 2 to 1) were generated.
3. Twenty random frequencies ranging from 200 to 300, and twenty ranging from 80 to 100 were generated.
4. The frequencies and ratios were distributed at random to generate twenty sets of four numbers.
5. Ten of those sets were, at random, assigned to congruent, and ten to incongruent trials.
6. For congruent trials, the higher ratio and higher frequency were used to generate the accompanying frequency of negative outcomes for the correct option (for example, a 2.9 to 1 ratio and a frequency of 256 positive outcomes produced a frequency of 88) - the smaller ratios and smaller frequency were used to generate the frequency of negative outcomes for the incorrect option.
7. For incongruent trials, the higher ratio and *smaller* frequency were used to generate the frequency of negative outcomes in the correct option while the lower ratio and *higher* frequency were used to generate the frequency of negative outcomes for the incorrect option.