EXAMINING THE ROLE OF FEELING OF RIGHTNESS WITH ANCHORING AND NUMBER OF MODELS

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

In Partial Fulfillment of the Requirements

For the Degree of Master of Arts

In the Department of Psychology

University of Saskatchewan

Saskatoon

By

Selina Wang

© Copyright Selina Wang, December, 2019. All rights reserve

Permission to Use

In presenting this thesis/dissertation in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis/dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis/dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis/dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis/dissertation.

Disclaimer

Reference in this thesis/dissertation to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this thesis/dissertation in whole or part should be addressed to:

Head of Psychology Department
University of Saskatchewan
9 Campus Drive, 154 Arts
Saskatoon, Saskatchewan S7N 5A5
Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Abstract

Feeling of Rightness (FOR) is a metacognitive experience accompanying people's intuitive answers that predicts subsequent answer changes (Thompson, Prowse Turner, & Pennycook, 2011). Previous research suggested cues that influence FOR also affect the ease with which an answer comes to mind, namely answer fluency. An issue that remains to be addressed is whether answer fluency drives the effect of FOR on subsequent behaviours pertaining to answer changes. The goal of a series of four experiments was to examine the relationship between FOR, answer fluency, and people's reanswer choices. Reasoners (N = 64) in each experiment were asked to determine the validity of 32 syllogisms that consisted of two models, single-model and multiple-models. Each syllogism was randomly paired with a question containing either a high anchor value (80% or 90%) or a low anchor value (10% or 20%). Reasoners then provided a FOR rating on a scale from 0 to 100 along with their reanswer choice for the first two experiments. The last two experiments served as the control experiments for which we removed the FOR judgements. Results suggested that influencing FOR without affecting answer fluency had no effect on people's subsequent reanswer choices. That is, when answers came to mind slowly, FORs were lower and people were more likely to choose to reanswer the problems. Possible explanations and limitations were further discussed in the paper.

Acknowledgements

I would like to thank Dr. Valerie Thompson who provided support, insight and expert supervision that greatly assisted the research. I would also like to thank the members of my advisory committee, Dr. Jamie Campbell and Dr. Carla Krachun for their insightful feedback, and my external examiner Dr. Regan Schmidt, for his time dedicated to this thesis. Lastly, I would like to thank my colleagues from the Cognitive Science Lab for their useful feedback and support.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

According to researchers at Cornell University, we make 226.7 decisions about food alone every day (Wansink & Sobal, 2007). Among the substantial choices we make each day, why do we choose to reflect on some decisions over others? In a multiple-choice exam, why do we change some answers, but not others? The answer may be a metacognitive one. Metacognition refers to the processes that monitor people's ongoing thought activities and processes that control the allocation of their mental resources (Nelson, 1990). Its function is analogous to a working thermostat that passively measures room temperature and controls the initiation and termination of the furnace. This thesis addresses issues of metacognitive monitoring and control as they apply to reasoning, which is also denoted as meta-reasoning.

1.1 Meta-Reasoning

Ackerman and Thompson (2017) developed a meta-reasoning framework to account for the processes that monitor and control people's reasoning and problem-solving activities based on the metamemory literature. The concept Judgement of Learning (JOL) in metamemory measures people's estimate of how well they have learned particular information, which directly influences their subsequent study choices (Metcalfe & Finn, 2008; Son & Metcalfe, 2000). When asked to memorize word pairs, people would be less likely to restudy the ones that they believed they would certainly recall on a later test. Similarly, when reasoners are asked to solve a reasoning task, their solution is posited to contain the answer itself and a metacognitive experience that accompanies it, which is referred to as the Feeling of Rightness (FOR) (Thompson et al., 2011; Thompson, Evans, & Campbell, 2013). Analogous to JOL, FOR can influence people's subsequent behaviours such as rethinking time and answer changes. As described next, FOR has been investigated in experiments using the two-response paradigm.

1.2. Relationship Between FOR and Reanswer Behaviours

In the two-response reasoning paradigm, reasoners are asked to provide a quick intuitive answer to each reasoning problem after which they are given as much time as they need to solve the problem again (Shynkaruk & Thompson, 2006). Rethinking time is measured as the response time of the second answer, and an answer change occurs when people reflect on their initial answer and change it on their second response. Previous research has shown that higher FOR judgements were associated with less rethinking time and lower likelihood of an answer change on a variety of reasoning tasks including base-rate problems, syllogisms, Wason's selection task, and denominator neglect (Thompson et al., 2011; Thompson et al., 2013; Thompson & Johnson, 2014). In other words, if reasoners were confident that their answers were correct (i.e., high FOR), they would spend less time rethinking the problems and were less likely to change their original answers.

1.3. Predictors of FOR

Metacognitive judgements such as Judgements of Learning (JOLs) are not based on access to memory content, but the experiences associated with generating the item (Koriat, 2007). As such, they are based inferentially on cues, and are accurate only to the extent the cues are accurate. These cues include encoding fluency (Begg, Duft, Lalonde, Melnick, & Sanvito, 1989; Hertzog, Dunlosky, Robinson, & Kidder, 2003; Undorf & Erdelder, 2011), font size (Rhodes & Castel, 2008), and retrieval latency of relevant information (Benjamin, Bjork, & Schwartz, 1998).

Similarly, FOR is also cue-based and inferential, and these cues can affect FOR without influencing accuracy on the task (Bajšanski, Močibob, & Valerjev, 2014; Bajšanski, Zauhar, &

Valerjev, 2018; Prowse Turner & Thompson, 2009; Quayle & Ball, 2000; Shynkaruk & Thompson, 2006; Thompson et al., 2011). These studies have shown that a reasoner's answer confidence judgements and accuracy are not well aligned in many reasoning domains, as indicated by the low correlations between the two variables. These results suggest that people can feel that they are right even when they are wrong.

One cue that has been demonstrated to influence FOR is answer fluency, the ease with which an answer comes to mind. There are two ways to experimentally identify such cues: correlation and manipulation. These two methods differ in that the latter involves directly manipulating independent variables while measuring and collecting behaviours such as response time that are associated with reasoning, which allows the interpretation of causation. However, prior attempts to manipulate FOR as described below were indirect in that they also affected answer fluency (Thompson et al., 2011; Thompson et al., 2013). Thus, it is less clear whether FOR per se predicts people's subsequent choices or answer fluency is the key that drives FOR and people's choices.

Thompson and colleagues (2011) manipulated FOR using base-rate and syllogistic problems with the two-response paradigm. In a classic base-rate task, participants are given two pieces of information: the probability of an individual belonging to one of two groups which is referred to as the base rate, and a personal description that favours membership in one of the two groups. When the two pieces of information point to the same group, the problem is considered non-conflicting; it is conflicting otherwise. It was found that FORs were lower for conflicting problems than for their non-conflicting counterparts, but the latter was also more fluent than the former in terms of response time (RT). In addition, there were longer rethinking times and a greater probability of answer change for conflict than non-conflict problems. For each person,

Thompson et al. (2011) also took the median RT of the initial responses, and compared FORs for RTs greater than and less than the median. Answers that were fluently generated (less than the median RT) were given higher FORs than their less fluent counterparts. These data revealed that answer fluency affected FORs which in turn led to the downstream behaviours associated with rethinking time and answer change.

In two further studies, Thompson and colleagues (2011, 2013) manipulated the availability of heuristics as a way to influence FOR. On a syllogistic reasoning task, the min heuristic is a non-logical shortcut for determining the validity of conclusions by evaluating how informative the premises and conclusion are (Thompson et al., 2011). Additionally, in a modified version of the Wason selection task, participants were given rules in the form of conditional statements (e.g., if p then q), and cards with a letter on one side and a number on the other side (Thompson et al., 2013). Their task was to determine whether the rule was true or false by evaluating each card. Results showed that participants using a two-response paradigm processed matching trials more fluently than non-matching trials when the matching heuristic (i.e. choosing cards with names mentioned in the rule) was available for use (Thompson et al., 2013). Of note, however, is the finding that answers generated by the heuristics were more fluent and given higher FORs than those that were not, and rethinking time was shorter and answer changes were less frequent for these corresponding problems. The two prior studies showed people's subsequent reanswering behaviours were determined by cues affecting FOR as well as answer fluency; thus, it was difficult to tease apart the effects generated by FOR from answer fluency.

To summarize, previous research showed the role of FOR in predicting people's reanswering behaviours (Thompson et al., 2011; Thompson et al., 2013); however, correlations between answer fluency and FOR were consistently observed such that each variable that

affected FOR and the downstream behavioural effects (i.e., answer change) also affected fluency. Therefore, the question that remains to be answered is whether manipulating FOR per se would be sufficient to predict people's subsequent reanswering behaviours independently of answer fluency. It is important to answer this question because it would shed light on the mechanism of meta-reasoning. Therefore, the goal of the following experiments was to examine the role of FOR in relation to subsequent reanswering choices by using a cue that directly manipulated FOR, and a cue that indirectly affected FOR (through the effect of answer fluency). To this end, as is explained next, we employed an "anchoring" manipulation as a cue that could directly influence FOR without affecting fluency.

1.4. Anchoring

Anchoring occurs when people incorporate a previously encountered value into a subsequent estimate, even when that value is irrelevant to the estimate. That is, people would generally provide a higher estimate if they encounter a high initial number. In a demonstration conducted by Tversky and Kahneman (1974), participants judged whether the proportion of African nations in the UN was higher or lower than an arbitrary anchor. The anchor point was determined by spinning a wheel of fortune, which was witnessed by the participants. Participants who encountered a higher anchor (i.e., 65%) gave higher estimates than those who saw a lower anchor (i.e., 10%).

Several theories attempt to explain the cognitive mechanisms of the anchoring effect, but two popular theories are the selective-accessibility theory and the scale distortion theory. The selective-accessibility theory posits that information relevant to the anchor value are activated, which causes people to give estimates that are consistent with it (Chapman & Johnson, 1999; Strack & Mussweiler, 1997; Mussweiler & Strack, 2000). For example, when asked to estimate

the price of a car after encountering a high anchor, features that are associated with an expensive car are activated such as a powerful engine. As a result, people tend to estimate a higher price for the car. The scale distortion theory on the other hand provides an alternative account, suggesting that the anchoring effect is caused by the distortion of the psychological scale (Frederick & Mochon, 2012). Based on the contrast effect (e.g., a dark room is perceived darker after walking out from a bright room), a large number on a scale feels even larger when the anchor value is small. As a result, people are likely to adjust their scale by moving towards to the smaller number in order to compensate for the distortion. Although the underlying cognitive mechanisms of the anchoring effect are still under study, the anchoring effect is a robust phenomenon which has been extensively studied in persuasion, attitude, judgments and decision-making (for review, see Furnham & Boo, 2011). However, it has not been widely investigated as a potential cue to examine metacognitive judgments, but the available data suggest that anchoring can be used to manipulate metacognitive judgements such as JOL (England & Serra, 2012; Yang, Sun, & Shanks, 2017; Zhao, 2012; Zhao and Linderholm, 2011), and by extension, FOR.

Specifically, Zhao and Linderholm (2011) and Zhao (2012) explored the anchoring effect on metacomprehension, which is people's ability to judge their own understanding of text materials. Participants were given texts to study and were asked about how well they would perform on future tests for the materials they just studied. Prior to providing their judgements, anchor values in the form of information on past peer performance with the same study content was shown to the participants. Participants who received high anchors (95%) gave higher judgements for their performance compared to those who saw low anchors (55%). Zhao (2012) also examined the effect of anchoring on people's retrospective judgments. Participants were asked to evaluate how well they did on a comprehension test on a scale from 0 – 100% after

studying for and taking the test. Again, participants who saw high-anchor information before rating their comprehension made higher retrospective metacomprehension judgements than those who encountered low-anchor information. A similar study was conducted using peer performance information as anchor values in a task of studying paired-associate words, and the results were consistent with prior Zhao and colleagues' findings (England & Serra, 2012).

In these studies, the anchoring values were informative, in that they provided participants with relevant information about peer performance of the same task. When the anchor values were irrelevant to the task performance, namely uninformative anchors, the relationship between anchoring and the metacognitive judgements was less clear (England & Serra, 2012; Zhao, 2012; Zhao and Linderholm, 2011). To address this gap in the research and to elucidate the role of anchoring in metamemory monitoring and control, Yang et al. (2017) conducted a set of experiments. In one of their experiments, participants studied a weakly-associated word-pair and were then told to answer the question "Is the likelihood you would be able to remember the preceding word pair in 5 min higher or lower than [10%/20%/30%/70%/80%/90%]?" In contrast to providing information about past peer performance, the anchoring information in this case presumably had no relevance to performing the task. Low anchors were comprised of 10%, 20%, and 30%, whereas high anchors were made up of 70%, 80%, and 90%. Each anchor value was randomly assigned to the word-pair, and each anchor value appeared equally often. Participants then provided a JOL score from 0 to 100, indicating the probability that they would be able to remember the pair in 5 min. Although the actual recall performance was not different between the high-anchor and low-anchor condition, JOLs were rated higher on high-anchor word-pairs compared to low-anchor counterparts. In the fourth experiment, participants were instructed to provide their restudy choices after making each JOL. More specifically, the participants

indicated whether they would like to study the previous word-pair again after they had seen all the word-pairs, although, in reality, the word pairs were not presented to them the second time. Consistent with Yang et al.'s (2017) previous findings, JOLs for the high-anchor word-pairs were higher than for low-anchor word-pairs, and participants chose fewer high-anchor pairs for restudy than their low-anchor counterparts. These results indicated that anchoring can produce a downstream effect on participants. In light of prior research on the anchoring effect in the restudy choices, we used anchor values in our experiments in order to directly influence people's FOR judgements of the task. We reasoned that anchor values were shown after participants provided their answers; thus, the anchoring effect would not affect answer fluency. Additionally, in order to test whether variables that affect FOR without affecting fluency would also influence reanswer choices, the reasoning task used in our experiments was syllogisms.

1.5. Models in Syllogistic Reasoning

Syllogisms represent a form of deductive reasoning. Each syllogism is made up of three statements, which include two premises and a conclusion. The conclusion of each syllogism contains two terms (e.g., A and C) presented in each premise, and a B term is always repeated in the premises. The reasoning task used in our experiments required participants to judge the validity of the syllogism's conclusion. Here is an example of one syllogism:

All of A are B.

All of B are C.

Therefore, all of A are C.

The mental model account of syllogistic reasoning posits that people start by constructing a single mental model that represents the relationships denoted by the premises (Johnson-Laird & Byrne, 1991). People subsequently derive a conclusion that is consistent with the initial model. In the case where a conclusion is presented for evaluation of its validity (i.e., does it follow logically from the premises), people test if the conclusion is consistent with the model. They should then need to test the conclusion against the possible alternative models of the premises, although people often neglect this step and end up accepting invalid conclusions. If it is not consistent with any other possible model, the conclusion is rejected. A valid conclusion is one that logically follows and is necessitated by the premises given; an invalid conclusion may be consistent with some of the possible models, but is not necessitated by the premises.

The number of models that can be used to represent the premises is related to the difficulty of the problem, which was a variable that was manipulated in the following experiments. We assumed we would be able to manipulate answer fluency by manipulating problem difficulty. Single-model problems require the construction of one model to determine the validity of the conclusion, whereas at least two models are needed for multiple-model ones. Examples of single-model and multiple-model syllogisms are shown below on the left and right respectively:

All of the dentists are painters.	None of the dentists are painters.
All of the painters are bicyclists.	All of the painters are bicyclists.
Therefore, some bicyclists are dentists.	Therefore, some bicyclists are not dentists.

Previous study found that people spent more time solving multiple-model syllogisms than their single-model counterparts (Copeland & Radvansky, 2004). Due to prior use of response time as a proxy measure for answer fluency (Thompson et al., 2011, 2013), we could use the difficulty variable (i.e., number of models required to deduce validity) to examine the effect of FOR on reanswer choices through the fluency effect.

In addition to slower response times, the accuracy of syllogistic reasoning tends to decrease as the number of models increases (Bara, Bucciarelli, & Johnson-Laird, 1995; Johnson-Laird & Bara, 1984; Klauer, Musch, & Naumer, 2000; Quayle & Ball, 2000). One explanation for this is that people often represent only one mental model, which is adequate for single-model problems, but multiple-model syllogisms require people to search for, represent (two or more mental models) and test alternative representations of the premises, which is more cognitively demanding (Ball, Phillips, Wade, & Quayle, 2006). Contrary to this research, however, Prowse-Turner and Thompson (2009) did not observe an accuracy difference between the two types, but single-model syllogisms were rated higher on confidence judgments than their multiple-model counterparts. Therefore, this study provided evidence that confidence judgement can be dissociated from accuracy. Other studies have also shown confidence as a poor indicator of accuracy in syllogistic reasoning (Bajšanski & Močibob, 2014; Quayle & Ball, 2000; Shynkaruk & Thompson, 2006; Thompson et al., 2011). Therefore, we predicted higher FORs for single-model and fluent problems than their counterparts independently of accuracy.

1.6. Summary

We attempted to investigate the role of FOR in predicting reanswer choices in syllogistic reasoning. To this end, we exploited a cue that was predicted to directly affect FOR, which was

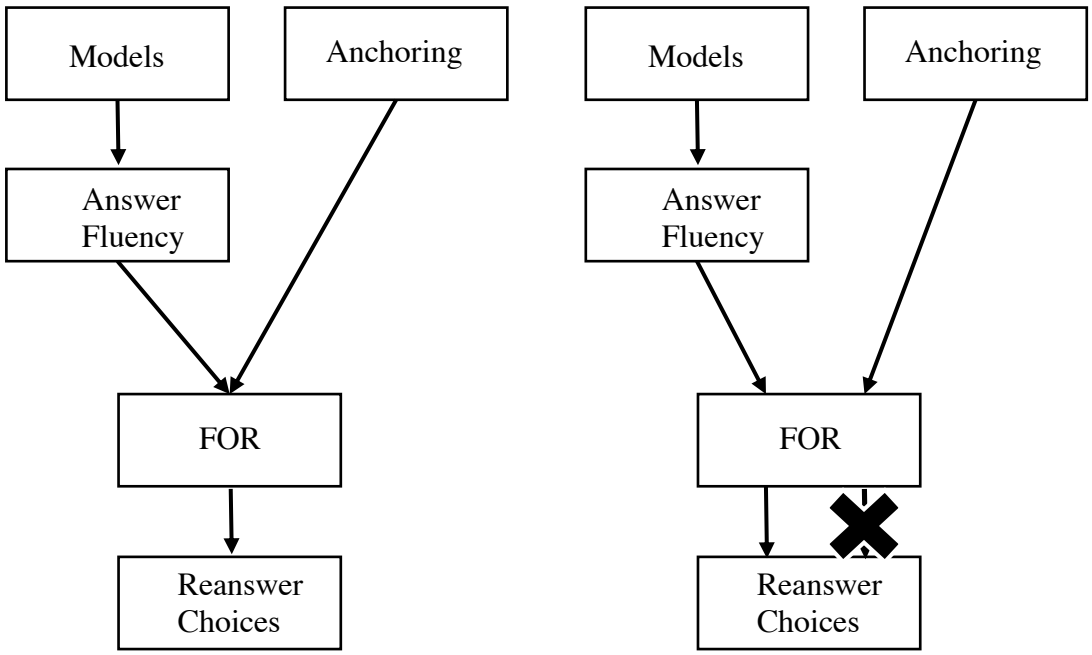
the size of anchors, and a cue that was predicted to influence FOR through the effect of answer fluency, which was the number of models.

In the first two experiments, we showed participants a random, uninformative anchor after they had solved each syllogism, and then asked them to give a FOR judgement based on their previous response. Given that participants provided their validity responses before seeing the anchor values, the effect of anchoring should not affect the participant's response time, which is a proxy measure of answer fluency. Participants then indicated their reanswer choice for the previous problem, that is, they indicated whether they would like to solve the preceding problem again in order to improve their overall score. In reality, they never solved the problems again. This measure of reanswer choices differed from previous studies, which examined people's actual behaviours of reconsideration and answer change (Thompson et al., 2011; Thompson et al., 2013), but was similar to the measures used in Yang et al.'s anchoring study (2017). We conducted two more experiments for which the FOR question was eliminated. The purpose of these experiments was to ensure participants' performance was not influenced by the act of providing their FOR ratings.

We formulated two alternative hypotheses regarding the relationships among answer fluency, FOR, and reanswer choices (see Figure 1.1). Hypothesis A: FOR can be predicted by anchor values and number of models (i.e., single- and multiple-model), which in turn, would affect people's reanswer choices as illustrated in Figure 1.1a. More specifically, we predicted that people would give higher FORs for high-anchor problems and for single-model syllogisms than their counterparts, and they would also be less likely to choose to reanswer these problems. Hypothesis B: only cues that affect the experience of answer fluency (i.e., number of models) would predict subsequent reanswer choices as depicted in Figure 1.1b. Those cues that do not

influence answer fluency may affect FOR, but they would not have any effects on reanswer choices. Therefore, answer fluency is the key factor for predicting reanswer choices. According to this hypothesis, we predicted that the number of models would influence answer fluency and FOR, which in turn would affect reanswer choices. On the other hand, anchoring would only affect FOR, but not reanswer choices.

All four experiments followed a within-subject design. Data were analyzed with a 2 (Anchor [low, high]) x 2 (Model [single, multiple]) x 2 (Validity [valid, invalid]) repeated-measures ANOVA.



(a) Hypothesis A (b) Hypothesis B

Figure 1.1 Flowcharts depicting two hypothesized paths of FORs in the current experiments.

Chapter 2. Experiment 1

The paradigm used in Experiment 1 closely matched the study conducted by Yang et al. (2017). We examined the effect of uninformative anchors on FORs and as well as reanswer choices. Reasoners were instructed to solve the syllogisms intuitively in order to be consistent with previous work on FOR (Thompson et al., 2011; Thompson et al., 2013). The responses collected were FORs, reading time (i.e., the time people spent reading the syllogisms), response time (i.e., the time from people finishing reading the questions to giving their responses), reanswer choices, and accuracy. To obtain a proxy measure of answer fluency, we summed response time and reading time.

2.1. Method

2.1.1. Participants

Sixty-four participants (35 females, 29 males, $M = 22$ years) were recruited from the University of Saskatchewan. They took part in the study for partial course credit.

2.1.2. Materials

The reasoning task was performed on a Microsoft Windows laptop computer with a 1920 x 1080 resolution display. Text instructions and stimuli were presented in black with an 18-point Courier New font, displayed on a white background.

Participants solved 32 syllogisms and 4 practice problems in the E-Prime 2.0 Software Tools program (Psychology Software Tools, Pittsburgh, PA). The reason we chose to include 32 stimuli was that this is a 2 (model) x 2 (anchor) x 2 (validity) repeated-measures within-subject design, and thus, each cell contained 4 items. Each of the syllogisms was comprised of 2 two-term premises (e.g., All of the “A” are “B”; All of the “B” are “C”) and a conclusion that related the “A” and “C” term (e.g., Therefore, some “C” are “A”). The A, B, and C terms all referred to

occupations (see Table 2.1 for examples). Among the syllogisms we used, 16 were from a published study (Prowse Turner & Thompson, 2009), and we created the remaining 16 syllogisms and the practice problems. To be consistent with the materials used in the previous study (Prowse Turner & Thompson, 2009), three moods (see Appendix A for details) were chosen for the single-model problems (AA, IA, and AI) and four moods were chosen for the multiple-model problems (AE, EA, IE, and EI). We also attempted to control for another factor in syllogisms called "figure" , which refers to the sequence in which the A, B and C terms are presented (See Appendix A for details). Participants received six Figure-1, four Figure-2, and six Figure-4 single-model syllogisms. Figure-3 single-model problems were not included because we attempted to be consistent with previous research (Prowse Turner & Thompson, 2009). Participants were also presented with five Figure-1, four Figure-2, two Figure-3, and five Figure-4 multiple-model problems. Additionally, the AC and CA conclusion orders were equally likely.

Table 2.1. Examples of syllogisms used in the experiment.

	Valid	**Invalid**
Single-model	All of the dentists are painters. All of the painters are bicyclists. Therefore, some bicyclists are dentists.	Some of the gardeners are psychologists. All of the gardeners are models. Therefore, all psychologists are models.
Multiple-model	None of the dentists are painters. All of the painters are bicyclists. Therefore, some bicyclists are not dentists.	All of the gardeners are psychologists. None of the models are gardeners. Therefore, some psychologists are models.

The validity of the syllogisms was manipulated such that half of the syllogisms were valid and the other half were invalid. Assuming all premises were true, valid syllogisms were those that necessarily followed from the premises. Invalid syllogisms were possibly true given

the premises[1], but were not necessitated by them (see Table 2.1 for examples). To control for any possible effect of content on validity, we counterbalanced the content of the syllogisms and created four lists to ensure that the occupation-related content in each premise pair was accompanied by two valid and two invalid conclusions.

The number of models was another variable we manipulated in the experiment. Sixteen problems were single-model, and the remaining 16 were multiple-model. Again, single-model syllogisms require the construction of one mental model to determine the validity of the conclusions, whereas at least two models are needed for multiple-model syllogisms.

Uninformative anchors were presented to the participants in the form of this question: "If you see the previous problem again, is the likelihood you would be able to solve it correctly higher or lower than [Anchor]%?" The anchor values were made up of low numbers (10 and 20) and high numbers (80 and 90). Each anchor value was randomly assigned to a syllogism. In addition, each type of syllogism was paired with an equal number of low and high anchors. For example, valid single-model syllogisms were accompanied by two of 10%, two of 20%, two of 80%, and two of 90% anchors.

2.1.3. Procedure

Participants were group-tested with an experimenter present. They were given brief instructions about the experiment and were told to solve the reasoning problems with their intuition. To familiarize themselves with the task procedure, participants began with 4 practice problems. The order of the problems was randomized, and they were presented on the screen one at a time. The event sequence is displayed in Figure 2.1. On each trial, participants saw two

[1] Half of the invalid problems are often falsely endorsed as valid because they are consistent with the first model people generate (Evans et al., 1999). People tend to accept the conclusions from these invalid problems rather than rejecting them, even though the latter is the correct response.

premises and a conclusion with a dashed line in the middle. Once they read the syllogism, participants pressed the space bar to continue. The interval between the onset of the problem to pressing the space bar was recorded as the reading time. After pressing the space bar, the question that pertained to the validity of the conclusion appeared directly below the problem. Participants chose 1 on the keyboard if they thought the conclusion followed logically from the premises; they chose 3 otherwise. The time required to do so was marked as their response time. After that, they answered the following question about their reanswer choices: “If you see the previous problem again, is the likelihood you would be able to solve it correctly higher or lower than [Anchor]%?” Participants chose either higher or lower for this question. They then provided their FOR ratings from 0 to 100, indicating how right they felt about their previous answer. Participants were asked about whether they would like to solve this problem again after they have solved all the problems to improve their overall score. In fact, no problems were presented to them the second time. There was a manipulation check in the end, asking participants if they actually answered with their intuition.

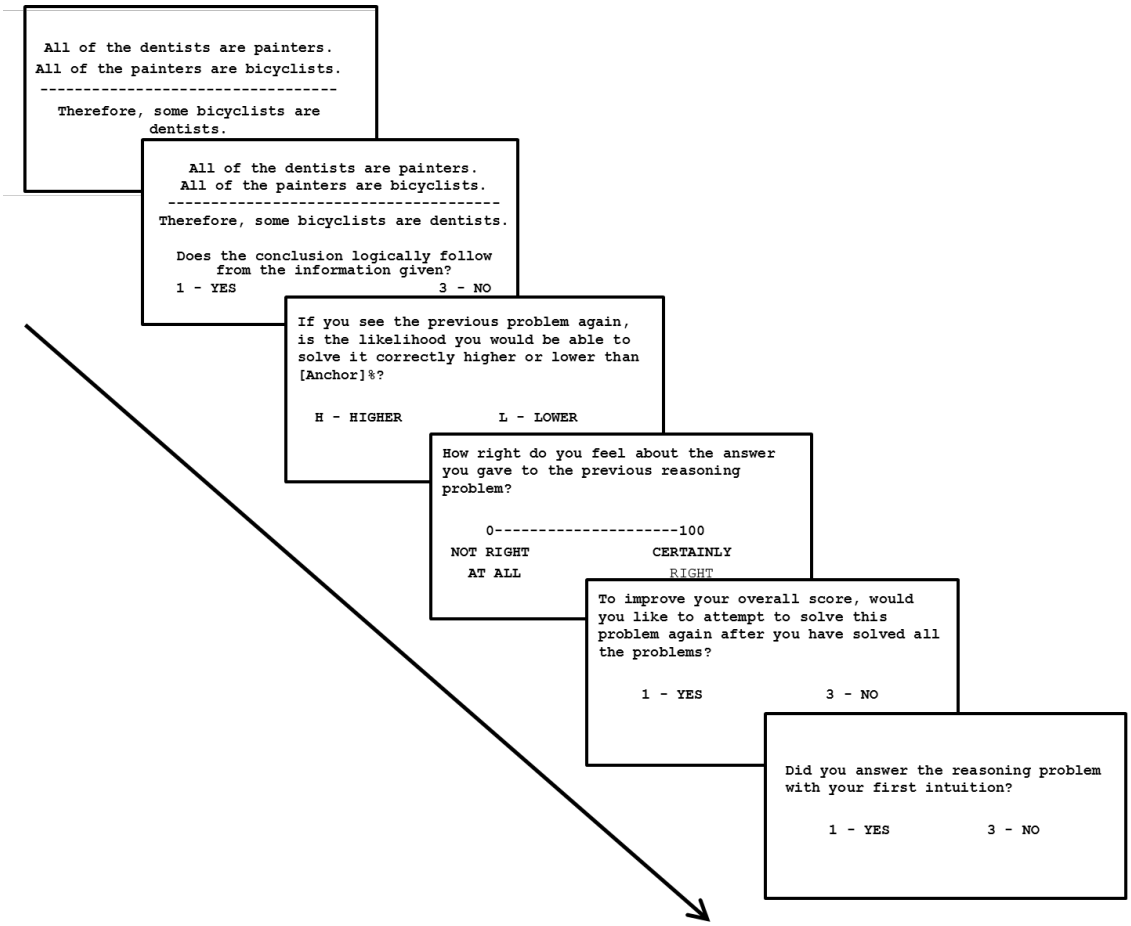


Figure 2.1. The trial progression for Experiment 1.

2.2. Results

Trials with missing FORs (i.e., the enter key was pressed without a numerical value) were discarded. Additionally, trials on which participants reported that they failed to provide an intuitive answer were also excluded from further analyses, which was about 4.8% of the data[2]. A 2 (Anchor [low, high]) x 2 (Model [single, multiple]) x 2 (Validity [valid, invalid]) repeated-measures ANOVA was performed on 4 dependent variables: FORs, reanswer choices, composite

[2] We also examined the effect of outliers on the data for all of the experiments. Removing RTs that were longer than 2 standard deviations away from the mean RT for each participant resulted in no significant changes. Therefore, we proceeded with the analysis without excluding the outliers.

RT (the sum of reading time and response time), and accuracy. Results with $p < 0.05$ were reported as significant. Paired t-tests were used to examine the simple main effects of the interactions.

2.2.1. FOR

The grand mean FOR rating collapsing across all levels of all factors was 82.71. The mean FOR in each of the eight cells in the 2 x 2 x 2 design are plotted in Figure 2.2. Consistent with our hypotheses, syllogisms paired with low anchors were rated lower on FOR (M = 81.61, SD = 1.41) than high-anchor syllogisms (M = 83.81, SD = 1.32), $F(1,63) = 7.559$, $p = .008$, $\eta_p^2 = 0.107$. As predicted, FORs were higher for single-model syllogisms (M = 84.92, SD = 1.28) than for multiple-model syllogisms (M = 80.51, SD = 1.45), $F(1,63) = 30.687$, $p < .001$, $\eta_p^2 = 0.328$. There was also a significant interaction between model and validity, $F(1,63) = 7.721$, $p = .007$, $\eta_p^2 = 0.109$. Values for the interaction are presented in Table 2.2. People gave higher FORs for single-model than for multiple-model syllogisms when the problems were valid (+5.88; $t(63) = 5.019$, $p < 0.001$), but this difference was smaller when the problems were invalid (+3.00; $t(63) = 4.290$, $p < 0.001$).

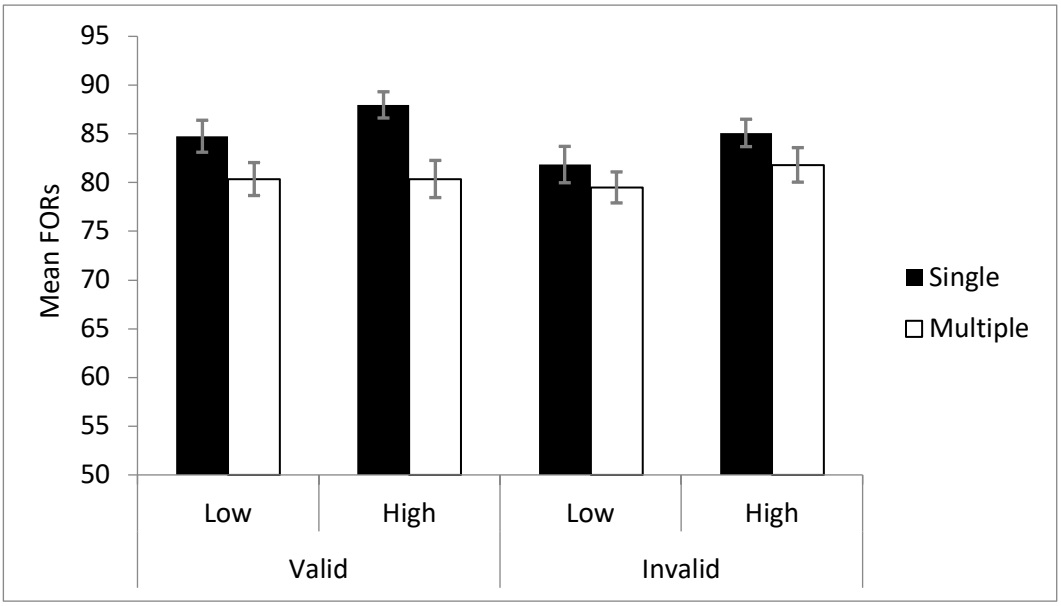


Figure 2.2. Mean FORs in Experiment 1 as a function of anchor, model and validity. Error bars represent standard errors.

Table 2.2. Mean FORs by model and validity.

Model	**Validity**	**Mean**	**Std. Error**	**N**
Single	Valid	86.36	1.26	64
	Invalid	83.47	1.48	64
Multiple	Valid	80.36	1.59	64
	Invalid	80.66	1.49	64

2.2.2. Reanswer choices

The overall mean probability of reanswering was 0.21. The data are plotted in Figure 2.3. Consistent with Hypothesis A and B, participants were less likely to reanswer single-model syllogisms (M = 0.18, SD = 0.03) than multiple-model syllogisms (M = 0.23, SD = 0.04), $F(1,63) = 8.845$, $p = .004$, $\eta_p^2 = 0.123$. The interaction between model and validity was also significant, $F(1,63) = 5.010$, $p = .029$, $\eta_p^2 = 0.074$. Values for the interaction are presented in Table 2.3.

When the problems were invalid, people were more likely to reanswer multiple-model syllogisms than single-model ones (+0.090; $t(63) = 4.797$, $p < 0.001$), but this difference was not found in valid problems (+0.015; $t(63) = 0.581$, $p = 0.564$). Contrary to Hypothesis A, the main effect of anchor was not significant, $F(1,63) = 0.007$, $p = .933$, $\eta_p^2 < 0.001$.

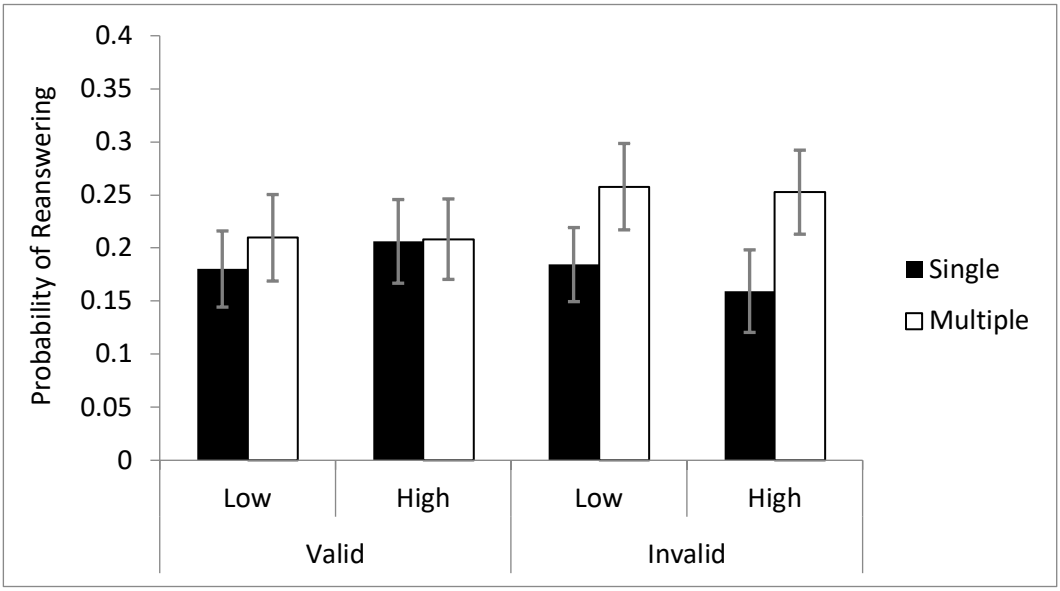


Figure 2.3. Probability of reanswering in Experiment 1 as a function of anchor, model and validity. Error bars represent standard errors.

Table 2.3. Probability of reanswering by model and validity.

Model	**Validity**	**Mean**	**Std. Error**	**N**
Single	Valid	0.193	0.035	64
	Invalid	0.172	0.033	64
Multiple	Valid	0.209	0.036	64
	Invalid	0.255	0.037	64

2.2.3. Composite RT

We combined the reading time and the response time in order to calculate the composite RT, which served as our proxy for fluency. Twenty-three participants' reading time data were not logged, and therefore, we performed the analysis with 41 participants. The data are presented in Figure 5. Participants solved single-model syllogisms faster ($M = 14.49$, $SD = 0.79$) than multiple-model syllogisms ($M = 16.70$, $SD = 0.96$), $F(1,40) = 13.955$, $p = .001$, $\eta_p^2 = 0.259$. Furthermore, the model and validity interaction was marginally significant, $F(1,40) = 3.833$, $p = .057$, $\eta_p^2 = 0.087$. Values for the interaction are displayed in Table 4. For valid syllogisms, participants solved single-model problems faster than multiple-model ones (-2.93; $t(40) = -3.380$, $p = 0.002$), whereas the difference between composite RT was marginally significant for invalid syllogisms (-1.27; $t(40) = -1.863$, $p = 0.07$). The size of the anchors had no effect on composite RT, $F(1, 40) = 1.266$, $p = .267$, $\eta_p^2 = 0.031$.

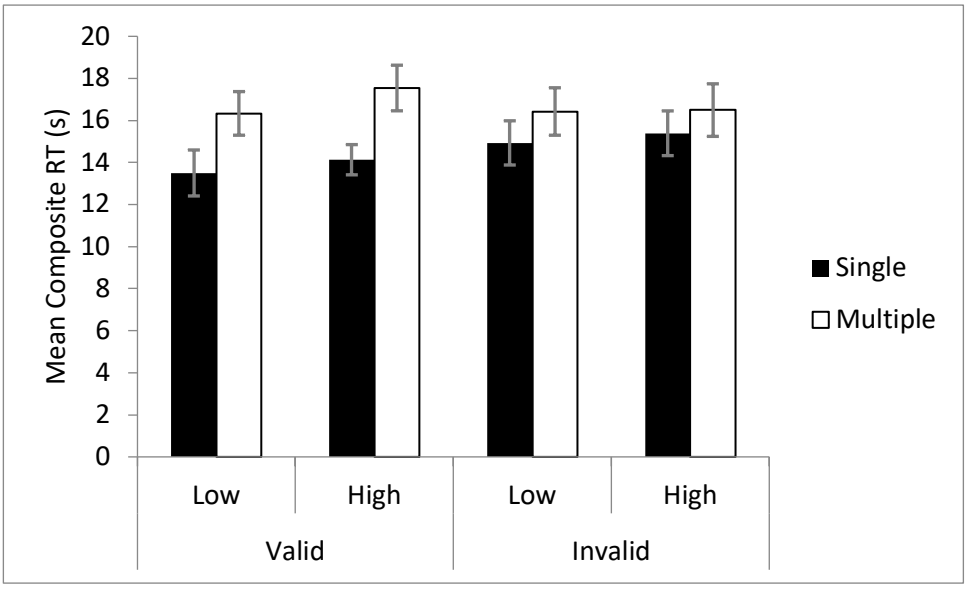


Figure 2.4. Mean reading time for Experiment 1 as a function of anchor, model and validity. Error bars represent standard errors.

Table 2.4. Mean reading time by model and validity

Model	Validity	Mean	Std. Error	N
Single	Valid	13.82	0.79	41
	Invalid	15.16	0.97	41
Multiple	Valid	16.94	0.99	41
	Invalid	16.46	1.08	41

2.2.4. Accuracy

The overall mean accuracy was 0.62. Data are plotted in Figure 2.5. Mean accuracy for valid syllogisms was higher ($M = 0.71$, $SD = 0.02$) than for their invalid counterparts ($M = 0.53$, $SD = 0.02$), $F(1,63) = 36.887$, $p < .001$, $\eta_p^2 = 0.369$. There was a significant interaction between model and validity, $F(1,63) = 10.195$, $p = 0.002$, $\eta_p^2 = 0.139$. These values are presented in Table 2.5. When the syllogisms were valid, participants were more accurate for single-model than for multiple-model problems (+0.09; $t(63) = 2.846$, $p = 0.006$), but this difference was absent for invalid syllogisms (-0.04; $t(63) = -1.309$, $p = 0.195$). The anchoring effect on accuracy was non-significant, $F(1,63) = 0.610$, $p = 0.438$, $\eta_p^2 = 0.010$.

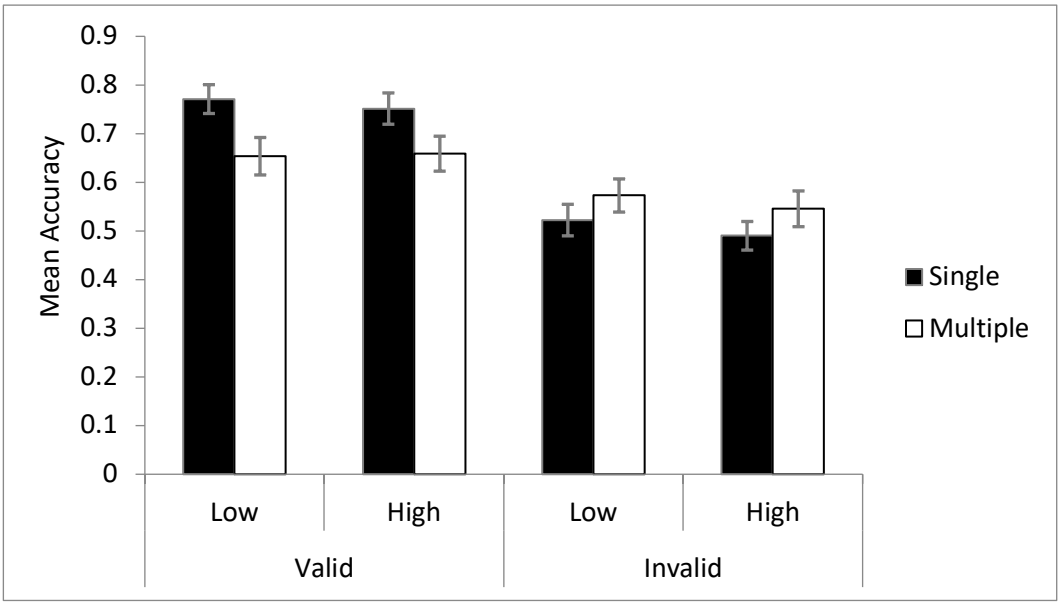


Figure 2.5. Mean accuracy for Experiment 1 as a function of anchor, model and validity. Error bars represent standard errors.

Table 2.5. Mean accuracy by model and validity of syllogisms

Model	**Validity**	**Mean**	**Std. Error**	**N**
Single	Valid	0.76	0.03	64
	Invalid	0.51	0.02	64
Multiple	Valid	0.66	0.03	64
	Invalid	0.56	0.03	64

2.2.5. Fluency Defined by Item RT

To examine the relationship between fluency, re-answer choices, and FOR, we computed a median composite RT (i.e., the sum of reading time and response time) for each participant[3].

[3] We also computed a median reading RT for each of the 41 participants whose reading time was logged. Consistent with the composite RT analysis, fluently read problems were given higher FORs ($M = 85.95$, $sd = 11.00$) than their less fluent counterparts ($M = 81.12$, $sd = 10.82$), $t(40) = 5.963$, $p < 0.001$. The effect of fluency on re-answer choices was only marginally significant, $t(40) = -1.748$, $p = 0.088$. However, the trend was similar to the composite RT data,

Then, we divided their responses into those that were fluently and disfluently generated. Fluently generated answers had composite RTs shorter than the participants' median composite RT and disfluent items had longer composite RTs. Consistent with previous research (Thompson et al., 2011; Thompson et al., 2013), fluently produced answers were given higher FORs ($M = 83.54$, $sd = 10.58$) than their disfluent counterparts ($M = 81.11$, $sd = 11.13$), $t(40) = 4.937$, $p < 0.001$. Additionally, we compared FORs for items based on participants' re-answer choices. FORs for the items participants preferred to re-answer ($M = 75.47$, $sd = 15.87$) were lower than for those they did not prefer to re-answer ($M = 83.58$, $sd = 10.92$), $t(46) = -3.083$, $p = 0.003$. We also examined the fluency effect on re-answer choices, which was also significant, $t(40) = -1.998$, $p = 0.05$. Participants on average preferred reattempting disfluent problems ($M = 0.24$, $sd = 0.26$) than fluent ones ($M = 0.21$, $sd = 0.27$). These analyses suggested that fluently generated responses were associated with higher FORs which also lowered participants' likelihood of reattempting the problems.

2.3. Discussion

2.3.1. Number of Models

In the current experiment, we verified that the number of models affected answer fluency. That is, people were more fluent at solving single-model syllogisms than multiple-model ones. We further observed that people's FOR ratings were higher for single-model syllogisms than for their multiple-model counterparts, and subsequently, they were less likely to choose the former to reanswer. These results were consistent with either of our hypotheses, because both hypotheses support that FOR can predict reanswer choices when it is influenced by answer

in that people were more likely to re-answer disfluent problems ($M = 0.23$, $sd = 0.29$) than their fluent counterparts ($M = 0.19$, $sd = 0.26$).

fluency. Evidence from item-based RT analysis also demonstrated the expected relationship between FORs and reanswer choices in that FORs were lower for the problems people were willing to reattempt. However, the results on the size of anchors provided us with a clearer direction.

2.3.2. Size of Anchors

We were able to demonstrate the anchoring effect on FOR (Figure 2.2) without affecting fluency as shown by the non-significant results on reading time and response time. People gave higher FORs to problems paired with high anchor values than their low counterparts without affecting answer fluency. However, their reanswer choices were unaffected by the size of the anchors. These data exclusively supported Hypothesis B because it seems that only cues that influence FORs through the effect of answer fluency can predict subsequent reanswer choices.

One possible explanation is that the size of anchors, unlike the number of models, can affect the judgement of the experience (i.e., FOR), but not the experience per se. FOR is intended to capture the judgement of the experience of being correct, which like other judgements can be altered by means such as anchoring (England & Serra, 2012; Yang et al., 2018; Zhao, 2012; Zhao and Linderholm, 2011). It is plausible that answer fluency is the source of the actual experience, which in turn predicts people's subsequent reanswer choices. On the other hand, cues that directly affect FOR can only influence the judgement of the FOR, but the actual experience that is captured by the FOR is unaffected. Based on our item-based RT analysis, FORs were lower for less fluent problems, and people tended to reattempt these problems more often than their fluent counterparts. In other words, fluency contributed to the sense of being right as reflected by the FOR rating, which in turn predicted people's succeeding reanswer choices. Thus, affecting FOR without affecting fluency may remove its behavioural consequences. In the

current experiment, the number of models affected fluency which also affected FOR as the responses for single-model syllogisms was processed more fluently than multiple-model ones, and people were less likely to reanswer the former. In contrast, although FOR was influenced by the anchoring effect, the difference in RTs as well as reanswer choices were similar for the low- and high-anchor syllogisms. The size of anchors served as an example of affecting FOR without affecting fluency, and as a result, people's subsequent choices were unaffected.

An alternative explanation is that the anchoring effect on reanswer choices was undetectable due to the task and measure we deployed in the experiment. These syllogisms were abstract and difficult compared to the task used in previous metamemory research, which was memorizing word-pairs (Yang et al., 2018). The difficulty of the reasoning task might have reduced people's overall motivation to reattempt the problems. As a result, the probability of people attempting to reanswer in the current experiment was only about 20% compared to the likelihood of restudying in previous research which was about 36%. People's reanswer choices might have resulted in a floor effect where the variances produced by the size of anchors could not be measured with the current task.

Another possibility is that FOR does not reliably predict people's intention to reanswer. In the current experiment, participants indicated whether they would like to solve each problem again, but in reality, they never solved any of the problems again. In order to be similar to Yang et al.'s study (2018), this behavioural measure was different than used in previous research, which examined participants' actual reanswering behaviours such as answer change and rethinking time (Shynkaruk & Thompson, 2006; Thompson et al., 2011; Thompson et al., 2013). Participants' intention might not reflect their actual behaviours; but this account is less likely because lower FORs were associated with higher reanswering probabilities within participants.

2.3.3. Accuracy

Cues such as the size of anchors affected people's FORs, but not their accuracy on the reasoning task, suggesting that confidence judgement and accuracy are dissociated. The data showed that the anchoring effect can directly bias people's confidence judgements, while leaving accuracy unaffected. Encountering a high anchor may lead us to report higher levels of confidence, but we are not necessarily correct. These results provide further support to the literature that confidence judgement in reasoning is poorly calibrated with accuracy (Bajšanski & Močibob, 2014; Prowse Turner & Thompson, 2009; Quayle & Ball, 2000; Shynkaruk & Thompson, 2006; Thompson et al., 2011).

Inconsistent with our expectation, the accuracy of single-model and multiple-model problems was similar, but this finding replicated the results found in Prowse Turner and Thompson's study (2009). Again, the single-model syllogisms used in the experiment contained several compelling lures such as the "all" word, which invited "all" in the conclusion. It is possible that people tended to construct an identity model for the quantifier "all" instead of forming a subset of models for it. They alternatively may simply match the quantifier of the conclusion with those presented in the premises which is also known as the matching heuristic to solve these problems, but in fact, employing this heuristic might be misleading (Wetherick & Gilhooly, 1995). For example, in the valid syllogisms below, participants might be tempted to infer the conclusion as invalid with the reasoning that the quantifiers "all" in the premises should be matched with an "all" quantifier in the conclusion.

All of the dentists are painters.

All of the painters are bicyclists.

Therefore, some bicyclists are dentists.

2.3.4. Validity

The results of the interaction between model and validity showed that people gave higher FORs to single-model syllogisms than multiple-model ones when the problems were valid, but this difference was not as profound with invalid problems. A similar interaction was present in accuracy. We did not anticipate this interaction to occur; therefore, we attempted to replicate this effect in the next experiment before making any conclusive inferences.

In the next experiment, the aim was to confirm the effects of number of models and size of anchors on reanswer choices. We attempted to replicate the current study again, but slightly changed the instructions by telling the participants "Some of the problems are very difficult." The reason for this was to reduce people's overall FORs with the hope of increasing the number of problems that people choose to reanswer.

Chapter 3. Experiment 2

The goal of Experiment 2 was to replicate the relationships amongst cues, FORs and reanswer choices as found in the previous experiment while attempting to make FOR judgements more varied by altering the instructions.

3.1. Method

3.1.1. Participants

Sixty-four participants (33 males and 31 females, M = 23 years) were recruited from the University of Saskatchewan. They took part in the study for course credit.

3.1.2. Materials and Procedure

The stimuli, design and procedure were the same as in Experiment 1, with one exception. The instructions of the current study emphasized that "Some of the problems are very difficult." This description was not present in the previous experiment.

3.2. Results

Trials with missing FORs and those that were not answered intuitively were excluded from further analyses, which accounted for 2.5 % of the data. A 2 (Anchor [low, high]) x 2 (Model [single, multiple]) x 2 (Validity [valid, invalid]) repeated-measures ANOVA was performed on 4 dependent variables: FOR, reanswer choices, composite RT (the sum of reading time and response time), and accuracy. Results with $p < 0.05$ were reported as significant. Paired t-tests were employed to reveal the simple main effects for significant interactions.

3.2.1. FOR

The mean FOR rating collapsing all levels was 83.82. The FOR data are plotted in Figure 3.1. As found in Experiment 1, participants gave higher FORs for single-model syllogisms (M = 85.90, SD = 1.56) than for their multiple-model counterparts (M = 81.73, SD =

1.73), $F(1,63) = 40.148$, $p < .001$, $\eta_p^2 = 0.389$. Once again, they also rated high-anchor syllogisms ($M = 85.09$, $SD = 1.53$) higher on FOR than low-anchor ones ($M = 82.54$, $SD = 1.80$), $F(1,63) = 8.238$, $p = 0.006$, $\eta_p^2 = 0.116$. The main effect of validity was significant, $F(1,63) = 7.623$, $p = 0.008$, $\eta_p^2 = 0.108$. Valid syllogisms ($M = 84.82$, $SD = 1.61$) were given higher FORs than invalid ones ($M = 82.82$, $SD = 1.70$). Consistent with Experiment 1, the interaction between model and validity was significant, $F(1,63) = 6.663$, $p = 0.012$, $\eta_p^2 = 0.096$. The values for the interaction are displayed in Table 3.1. To decompose this interaction, people gave higher FORs to single-model syllogisms than to multiple-model ones when the problems were valid (+6.13; $t(63) = 6.392$, $p < 0.001$), but the difference was smaller for invalid problems (+2.24; $t(63) = 2.187$, $p = 0.032$), replicating Experiment 1.

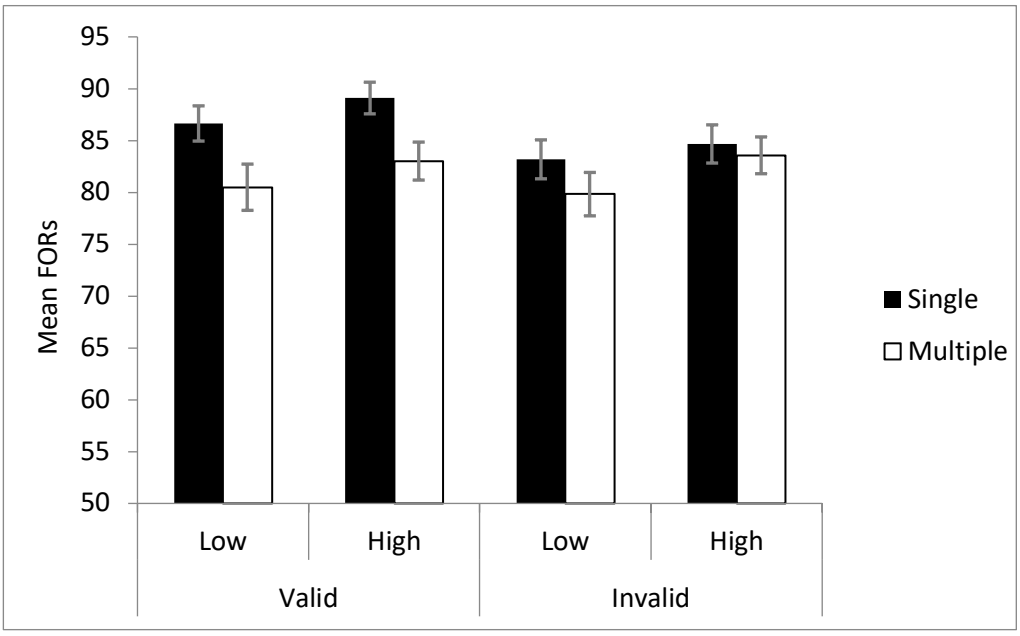


Figure 3.1. Mean FORs in Experiment 2 as a function of model, anchor, and validity. Error bars represent standard errors.

Table 3.1. Mean FORs by model and validity

Model	Validity	Mean	Std. Error	N
Single	Valid	87.88	1.48	63
	Invalid	83.93	1.71	63
Multiple	Valid	81.76	1.86	63
	Invalid	81.70	1.83	63

3.2.2. Reanswer Choices

The overall mean probability of reanswering was 0.20, only 0.01 lower than in Experiment 1. The data are illustrated in Figure 3.2. Contrary to Experiment 1, there was a marginally significant interaction between model and anchor, $F(1,63) = 3.805$, $p = 0.056$, $\eta_p^2 = 0.057$. Participants were more likely to reattempt the multiple-model problems than their single-model counterparts for the low-anchor problems (+0.05; $t(63) = 2.379$, $p = 0.020$), but the pattern disappeared for the high-anchor problems (+0.01, $t(63) = 0.267$, $p = 0.790$). Again, there was no anchoring effect on reanswer choices, $F(1,63) = 0.244$, $p = 0.623$, $\eta_p^2 = 0.004$.

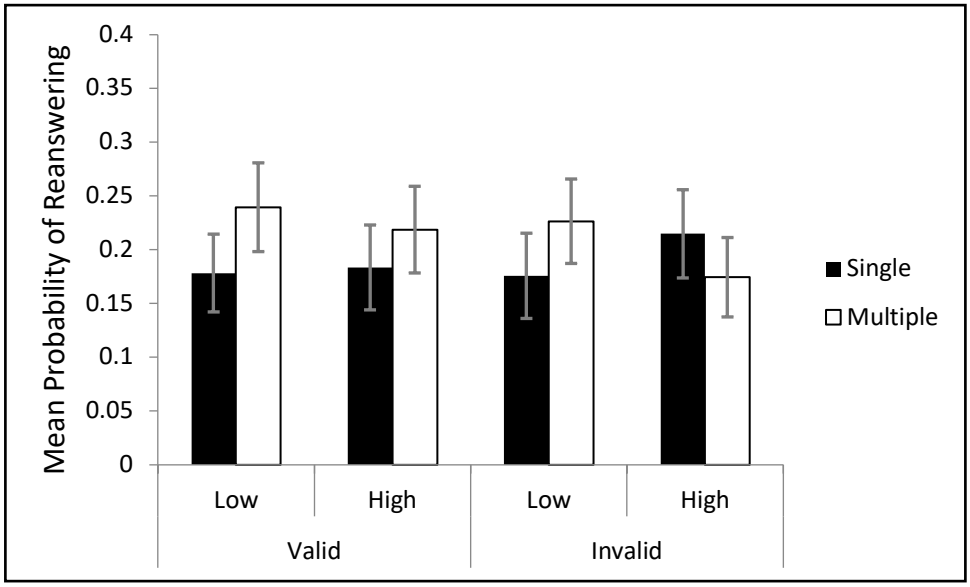


Figure 3.2. Mean probability of reanswering in Experiment 2 as a function of model, anchor, and validity. Error bars represent standard errors.

Table 3.2. Mean FORs by model and anchor

Model	**Anchor**	**Mean**	**Std. Error**	**N**
Single	Low	0.18	0.04	63
	High	0.20	0.04	63
Multiple	Low	0.23	0.04	63
	High	0.20	0.04	63

3.2.3. Composite RT

Experiment 1 showed that the analysis of composite RT produced similar results compared to the analysis of reading time alone, and there were no effects of any independent variables on response time. Therefore, we computed the composite RT, the combination of participants' reading time and response time to the problems. We proceeded with the analysis of composite RT instead of analyzing reading time and response time separately, attempting to

collect a more precise proxy measure of answer fluency. The overall mean composite RT was 20.02 seconds. The data are plotted in Figure 3.3. Replicating results found in Experiment 1, participants responded to single-model problems ($M = 18.40$, $SD = 0.99$) faster than multiple-model problems ($M = 21.64$, $SD = 1.36$), $F(1,63) = 24.920$, $p < .001$, $\eta_p^2 = 0.283$. Again, the main effect of anchor on RTs were non-significant, $F(1,63) = 0.376$, $p = 0.542$, $\eta_p^2 = 0.093$.

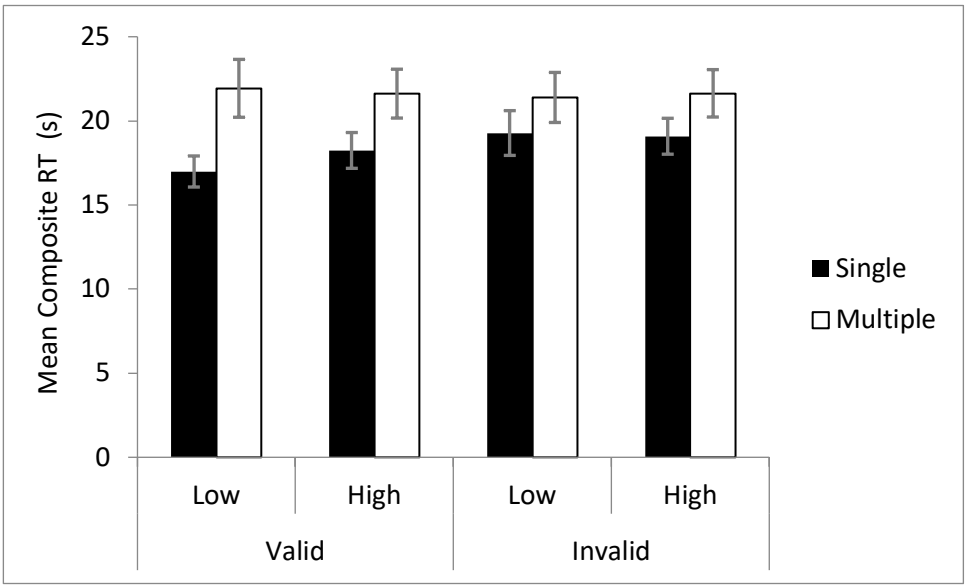


Figure 3.3. Mean composite RT in Experiment 2 as a function of model, anchor, and validity. Error bars represent standard errors.

3.2.4. Accuracy

The overall mean accuracy was 0.63. The data for accuracy are plotted in Figure 3.4. Contrary to Experiment 1, participants were more accurate on the single-model syllogisms ($M = 0.66$, $SD = 0.02$) than their multiple-model counterparts ($M = 0.59$, $SD = 0.02$), $F(1,63) = 13.805$, $p < .001$, $\eta_p^2 = 0.180$. There was also a main effect of validity, $F(1,63) = 12.716$, $p = .001$, $\eta_p^2 = 0.168$. The mean accuracy for valid syllogisms ($M = 0.68$, $SD = 0.02$) was higher than for invalid ones ($M = 0.58$, $SD = 0.02$). Consistent with Experiment 1, the interaction between model and validity was significant, $F(1,63) = 34.986$, $p < .001$, $\eta_p^2 = 0.357$. When the syllogisms were valid,

single-model problems were answered more accurately than multiple-model problems (+0.19, $t(63) = 7.994$, $p < 0.001$), but the difference was not present when the syllogisms were invalid (-0.04, $t(63) = -1.447$, $p = 0.153$). There was a marginally significant anchoring effect on accuracy[4], $F(1,63) = 3.849$, $p = .054$, $\eta_p^2 = 0.058$. The mean accuracy for single-model syllogisms was slightly higher than for multiple-model problems when paired with high anchors (+0.08, $t(63) = 2.533$, $p = 0.014$), but the difference was smaller for problems paired with low anchors (+0.06, $t(63) = 2.272$, $p = 0.026$).

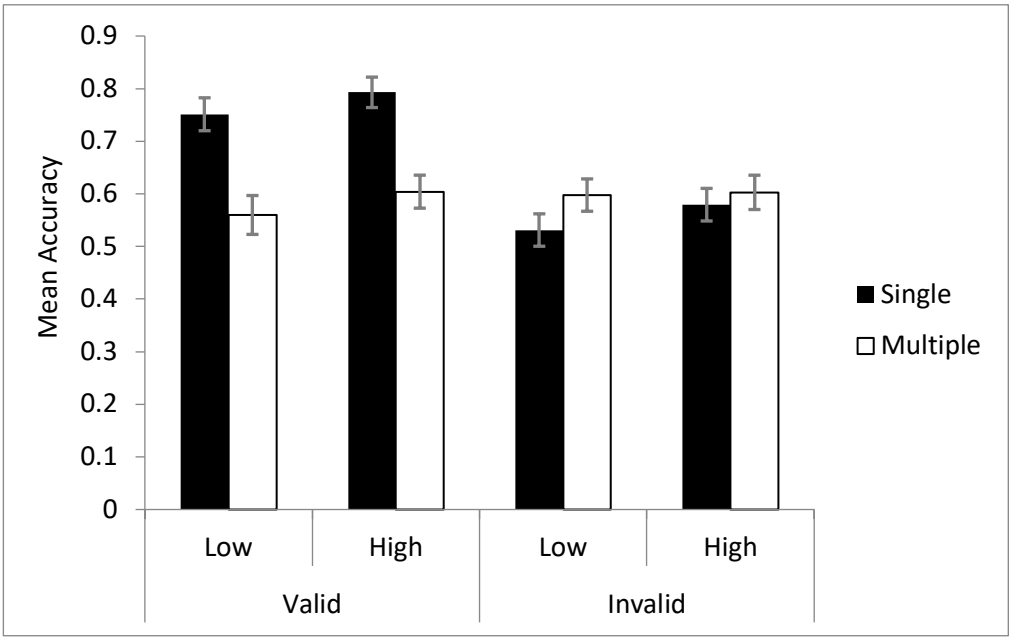


Figure 3.4. Mean accuracy in Experiment 2 as a function of model, anchor, and validity. Error bars represent standard errors.

[4] In Experiment 2, there was a marginally significant effect of the size of anchors on accuracy. This was likely to be a Type I error. The anchoring manipulation occurred after participants provided their responses, therefore, it was logically impossible that the anchors influenced participants' accuracy on the task.

Table 3.3. Mean accuracy by model and validity

Model	Validity	Mean	Std. Error	N
Single	Valid	0.77	0.02	63
	Invalid	0.56	0.02	63
Multiple	Valid	0.58	0.03	63
	Invalid	0.60	0.03	63

3.2.5. Fluency Defined by Item RT

As in Experiment 1, we computed a median composite RT for each participant. Composite RTs less than the median were coded as fluently generated and those that were longer than the median were considered as disfluent. Consistent with Experiment 1, we found that fluently generated responses were given higher FORs ($M = 85.57$, $sd = 13.53$) than their disfluent counterparts ($M = 82.26$, $sd = 13.33$), $t(63) = 3.560$, $p = 0.001$. We again compared the FORs of participants' responses according to their reanswer choices. The items participants preferred to reanswer were given lower FORs ($M = 74.46$, $sd = 21.26$) than those they were unwilling to reattempt ($M = 84.57$, $sd = 13.54$), $t(52) = -4.110$, $p < 0.001$. Additionally, there was a fluency effect on reanswer choices, $t(63) = -4.159$, $p < 0.001$. Participants were more willing to choose to reanswer disfluent problems ($M = 0.24$, $sd = 0.29$) than their fluent counterparts ($M = 0.16$, $sd = 0.26$).

3.3. Discussion

We attempted to remove the floor effect of reanswer choices in Experiment 1. In the current experiment, we modified the instructions by telling participants that "Some of the

problems are very difficult", but the instruction manipulation did not increase their reanswering probabilities.

3.3.1. Size of Anchors

Consistent with the results found in Experiment 1, the size of anchors influenced FORs without affecting answer fluency. People gave higher FORs to high-anchor syllogisms than their low-anchor counterparts, but their composite RT was not subject to the anchoring effect. Moreover, reanswer choices were not affected by the size of anchors. These data provided evidence supporting Hypothesis B, which posits that cues directly influencing FOR cannot predict subsequent reanswer choices unless they also affect answer fluency.

3.3.2. Number of Models

Replicating the effect of models on FORs from Experiment 1, the number of models affected people's FORs in that single-model syllogisms were rated higher on FORs than their multiple-model counterparts. The latter involves representing and testing two or more models, requiring more cognitive effort than the former. Furthermore, the number of models also influenced answer fluency. That is, people solved the single-model syllogisms more fluently than multiple-model ones, replicating the results found in Experiment 1. However, the results on reanswer choices were contrary to Experiment 1.

In the current experiment, one of the manipulated variables, number of models, failed to produce any effects on reanswer choices. We formed two hypotheses about the relationships among fluency, FOR, and reanswer choices, but both hypotheses suggest that cues (e.g., number of models) affecting fluency should influence FOR judgments, which in turn predict people's reanswer choices. Our finding pertaining to the number of models in the current experiment seemed to challenge the expected set of relationships described previously. It also appeared to

contradict the results from the item-based RT analysis. The divergence is described as follows: people responded to single-model syllogisms more quickly than multiple-model problems, but reanswer choices did not differ between the two types of models, even though the fluency analyses showed that people were more likely to reanswer the slow-responding problems on the individual level. Considering the relationship between fluency and reanswer choices from the item-based RT analysis, we would expect a higher likelihood of reanswering on the multiple-model problems since the composite RT for multiple-model syllogisms was longer (i.e., multiple-model syllogisms were less fluent) than single-model counterparts, but the data suggested otherwise. To foreshadow, we observed this null effect of number of models on reanswer choices in the next two experiments as well. A more detailed explanation for this divergence is provided in the summary section in Experiment 4, but the gist is that the relative answer fluency for each individual is the key predictor of reanswer choices.

3.3.3. Accuracy

In contrast to Experiment 1, the accuracy for single-model syllogisms was higher than multiple-model ones. This inconsistency might be due to sampling variability. It was possible that participants in the current experiment were less susceptible to compelling lures in the single-model syllogisms than participants from Experiment 1.

3.3.4. Validity

We also found that the participants gave higher FORs for valid problems than invalid ones, which was not found in Experiment 1. According to the one-model hypothesis (Evans, Handley, Harper, & Johnson-Laird, 1999), the reasoners construct one representation of the premises, if the conclusion is necessitated by the premises, the syllogism is judged to be valid; it is judged invalid otherwise. Invalid syllogisms were more difficult because the conclusion might

be consistent with the first model reasoners constructed, but it did not necessarily follow from the premises. Therefore, more than one representation of the premises was possibly required to evaluate invalid syllogisms. Reasoners might be aware that the current model would not be sufficient to determine the validity of the syllogism, but they were uncertain about how to proceed (Quayle & Ball, 2000). This experience is referred to as metacognitive uncertainty, which may have been reflected in our FOR measure. This may be the reason that lower FORs were given to the invalid syllogisms than their valid counterparts. Furthermore, the interaction between model and validity on FORs was also replicated in the current experiment as people gave higher FORs for single- than multiple-model syllogisms when the syllogisms were valid, but there was no significant difference between the two problem types for invalid problems. We do not have an obvious explanation for this result.

Chapter 4. Experiment 3

The third experiment served as a control experiment. The procedure followed the previous two experiments except that the question pertaining to FOR was eliminated. The concern was that incorporating FOR judgements may facilitate deliberate thinking, which would in turn lead to overstated effects like longer composite RT, increased accuracy, and a higher probability of reanswering. The goal of the following two experiments was to address this concern and verify that adding in the metacognitive question would not induce changes in people's behaviours. We hypothesized that the results on reanswer choices, composite RT, and accuracy in the current experiment would mirror the previous two experiments.

4.1. Method

4.1.1. Participants

Sixty-four participants (37 males and 27 females, $M = 21$ years) were recruited from the University of Saskatchewan. These participants took part in the study for partial course credit.

4.1.2. Materials and Procedure

The materials and procedure were the same as Experiment 1 and 2 with the exception of the FOR question. The participants in the current experiment did not see the FOR question. After they saw the question with the anchoring information, they then proceeded to indicate whether they would like to solve the previous problem again.

4.2. Results

According to participants' self-reports, responses that were not answered intuitively were discarded, which accounted for 3.7% of the data. A 2 (Anchor [low, high]) x 2 (Model [single, multiple]) x 2 (Validity [valid, invalid]) repeated-measures ANOVA was performed on 3 dependent variables: reanswer choices, composite RT, and accuracy. Results with $p < 0.05$ were

reported as significant. Paired t-tests were used to examine the simple main effects of significant interactions.

4.2.1. Reanswer Choices

The overall mean probability of reanswering was 0.20. Consistent with Experiment 2, the main effect of model on reanswer choices was non-significant, $F(1,63) = 0.169$, $p = 0.683$, $\eta_p^2 = 0.003$. Consistent with both Experiment 1 and 2, the main effect of anchor was also non-significant, $F(1,63) = 2.522$, $p = 0.117$, $\eta_p^2 = 0.038$.

4.2.2. Composite RT

The composite RT was computed exactly like the prior experiments. The overall mean composite RT was 15.74 seconds. The data are illustrated in Figure 4.1. Consistent with the previous experiments, we found a main effect of model on composite RTs, $F(1,63) = 44.777$, $p < 0.001$, $\eta_p^2 = 0.415$. Participants responded to the single-model syllogisms ($M = 14.48$, $sd = 0.66$) more quickly than their multiple-model counterparts ($M = 17.01$, $sd = 0.80$). However, there was also a main effect of validity, $F(1,63) = 6.771$, $p = 0.012$, $\eta_p^2 = 0.097$. The valid syllogisms were answered faster ($M = 15.27$, $sd = 0.67$) than the invalid ones ($M = 16.22$, $sd = 0.79$) by participants. Again, the main effect of anchor was non-significant, $F(1,63) = 0.008$, $p = .930$, $\eta_p^2 < 0.001$.

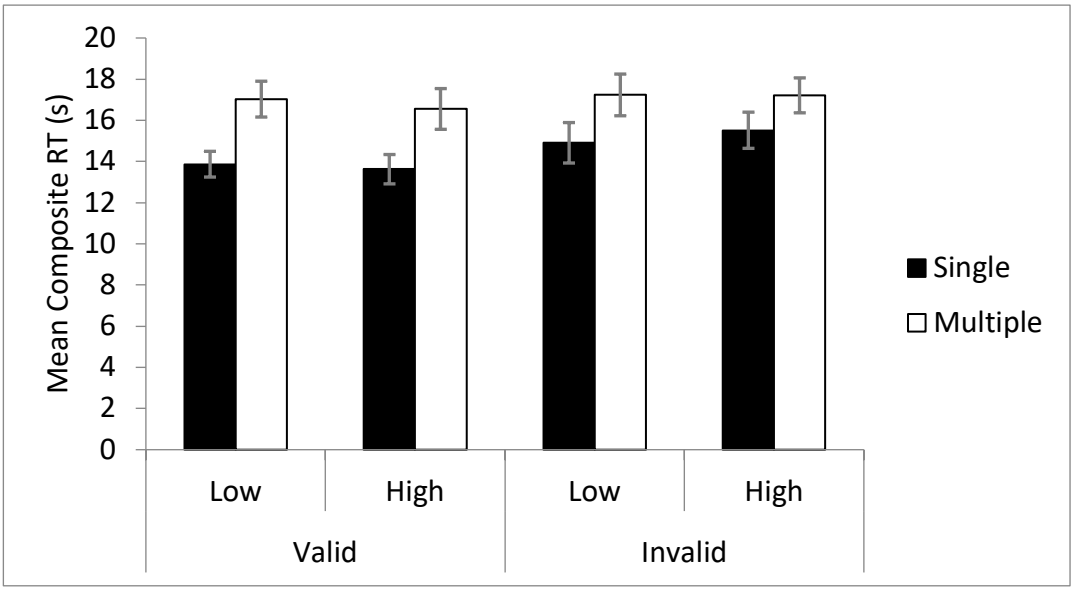


Figure 4.1. Mean composite RTs in Experiment 3 as a function of model, anchor, and validity. Error bars represent standard errors.

4.2.3. Accuracy

The overall mean accuracy was 0.61. Data are plotted in Figure 4.2. Consistent with Experiment 1 and 2, the accuracy for valid syllogisms ($M = 0.68$, $sd = 0.02$) was higher than for invalid problems ($M = 0.54$, $sd = 0.02$), $F(1,63) = 21.702$, $p < 0.001$, $\eta_p^2 = 0.256$. Similar to the previous experiments, the interaction between model and validity was also significant, $F(1,63) = 37.487$, $p < 0.001$, $\eta_p^2 = 0.373$. The values of the interaction are displayed in Table 4.1. When the problems were valid, participants were more accurate on the single-model syllogisms (+0.14, $t(63) = 5.391$, $p < 0.001$), whereas their accuracy was higher on the multiple-model syllogisms when the problems were invalid (-0.09, $t(63) = -2.871$, $p = 0.006$). Consistent with Experiment 1,

the main effect of model on accuracy was non-significant, $F(1,63) = 1.969$, $p = 0.165$, $\eta_p^2 = 0.03$.

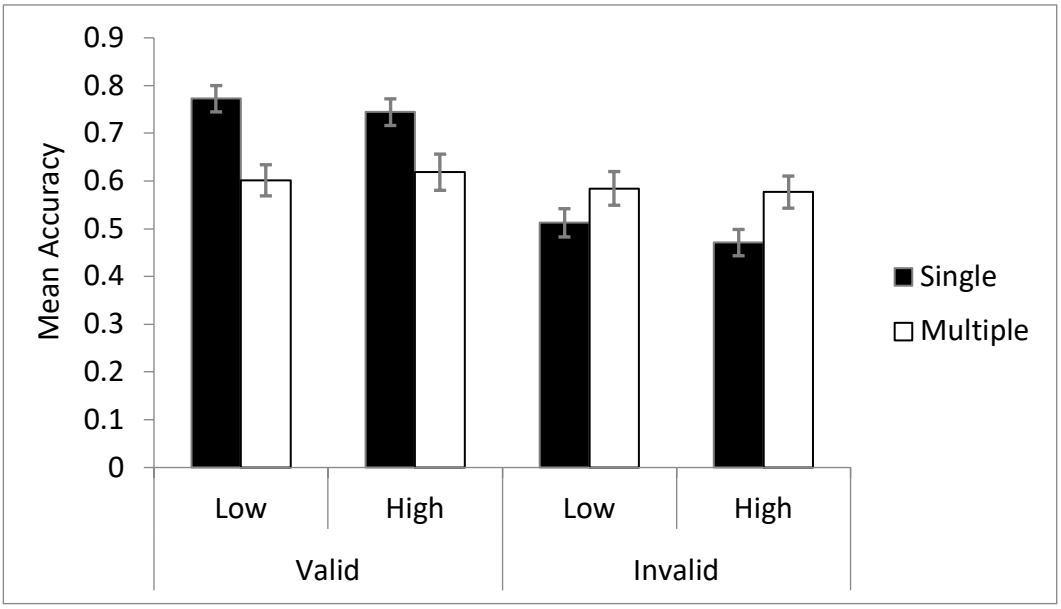


Figure 4.2. Mean accuracy in Experiment 3 as a function of model, anchor, and validity. Error bars represent standard errors.

Table 4.1. Mean accuracy by model and validity

Model	Validity	Mean	Std. Error	N
Single	Valid	0.76	0.02	63
	Invalid	0.49	0.02	63
Multiple	Valid	0.61	0.03	63
	Invalid	0.58	0.03	63

4.2.4. Fluency Defined by Item RT

For the item-RT analysis, we again calculated a median composite RT for every participant. Similar to findings in Experiment 1 and 2, people had a tendency to reattempt the problems that were answered less fluently ($M = 0.22$, $sd = 0.22$) than those that were answered more fluently ($M = 0.18$, $sd = 0.25$), $t(63) = -1.890$, $p = 0.063$.

4.3. Discussion

The results from the current experiment closely replicated Experiments 1 and 2, which are summarized in Tables 5.2 – 5.6 in the next chapter for comparison purposes. With respect to reanswer choices, number of models had no effect on people's preference for reattempting the problems, which was consistent with Experiment 2. This finding neither supported Hypothesis A nor Hypothesis B. Both hypotheses proposed that cues affecting answer fluency (e.g., number of models) would subsequently influence FORs, which in turn predict people's subsequent reanswer choices. However, the item-based RT analysis showed that people were more likely to reanswer the less fluent problems (i.e., multiple-model syllogisms according to composite RT data), even though overall they did not show preference for reattempting multiple-model over single-model syllogisms. The conundrum is further discussed in the next section.

Due to mixed results concerning the effect of models on reanswer choices in the previous two experiments, it was less clear whether incorporating FOR judgements had an impact on people's reanswer choices. The argument is that including FOR judgements may artificially influence people's subsequent reanswer choices, possibly alerting them to be more reflective on the succeeding task. However, evidence from the anchoring effect on reanswer choices hinted otherwise. The size of anchors exerted no effect on reanswer choices across three experiments. This consistent null effect of anchoring on reanswer choices suggested that incorporating FOR questions did not change people's reanswer choices compared to experiments without FORs. In other words, the size of anchors had no effect on reanswer choices regardless of the presence of FOR judgements. We needed to replicate these results found in Experiment 3, and that was the reason for conducting the fourth experiment.

Chapter 5. Experiment 4

In this fourth experiment, the goal was to replicate Experiment 3 in attempt to verify that including FOR judgements would not affect people's behaviours on the reasoning task.

5.1. Method

5.1.1. Participants

Sixty-four participants (36 males and 28 females, M = 28 years) were recruited from the bulletin board on the University of Saskatchewan website. They received $7.50 for their participation instead of receiving partial course credit as in the previous experiments.

5.1.2. Materials and Procedure

The stimuli, design and procedure were the same as Experiment 3.

5.2. Results

Based on participants' self-reports, deliberate responses were excluded, which was about 1.7% of the data. A 2 (Anchor [low, high]) x 2 (Model [single, multiple]) x 2 (Validity [valid, invalid]) repeated-measures ANOVA was performed on 3 dependent variables: reanswer choices, composite RT and accuracy. Results with $p < 0.05$ were reported as significant. Paired t-tests were employed to reveal the simple main effects of significant interactions.

5.2.1. Reanswer Choices

The overall mean probability of reanswering was 0.318. Consistent with Experiment 3, there was no main effect of model on reanswer choices, $F(1,63) = 0.002$, $p = 0.969$, $\eta_p^2 < 0.001$. The main effect of anchor was also non-significant, $F(1,63) = 2.622$, $p = 0.110$, $\eta_p^2 = 0.040$.

5.2.2. Composite RT

The overall mean composite RT was 17.52 seconds. Data are displayed in Figure 5.1. Again, we found a main effect of model on composite RTs, $F(1,63) = 38.140$, $p < 0.001$, $\eta_p^2 =$

0.377. Participants were faster at solving the single-model syllogisms ($M = 15.91$, $sd = 0.76$) than multiple-model ones ($M = 19.13$, $sd = 1.05$). Participants also responded to the valid problems ($M = 16.94$, $sd = 0.84$) more quickly than their invalid counterparts ($M = 18.10$, $sd = 0.97$), $F(1,63) = 6.001$, $p = 0.017$, $\eta_p^2 = 0.087$. The main effect of anchor was non-significant, $F(1,63) = 2.622$, $p = .110$, $\eta_p^2 = 0.040$.

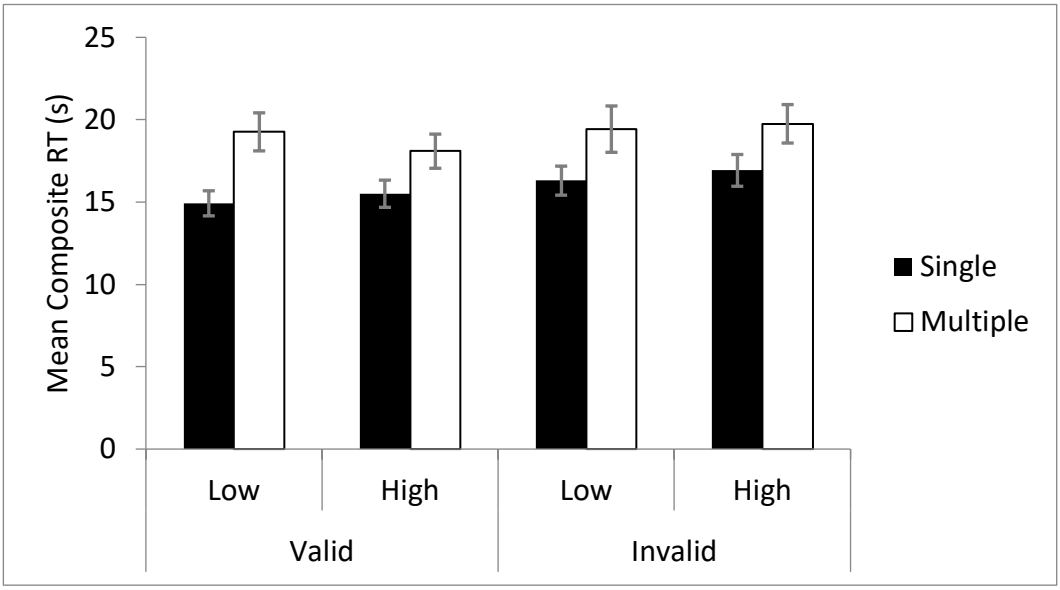


Figure 5.1. Mean composite RTs as a function of model, anchor, and validity. Error bars represent standard errors.

5.2.3. Accuracy

The overall mean accuracy was 0.66. Data are plotted in Figure 5.2. Consistent with previous experiments, there was a main effect of validity on accuracy, $F(1,63) = 29.807$, $p < 0.001$, $\eta_p^2 = 0.321$. Participants were more accurate on the valid syllogisms ($M = 0.74$, $sd = 0.02$) than their invalid counterparts ($M = 0.58$, $sd = 0.02$). Similar to previous experiments, the interaction between model and validity was also significant, $F(1,63) = 37.103$, $p < 0.001$, $\eta_p^2 = 0.371$. Participants correctly answered more single-model problems than multiple-model ones when the syllogisms were valid (+0.14, $t(63) = 5.325$, $p < 0.001$), but they were more accurate on

multiple-model problems than their single-model counterparts for the invalid problems (-0.09, $t(63) = -3.541, p = 0.001$).

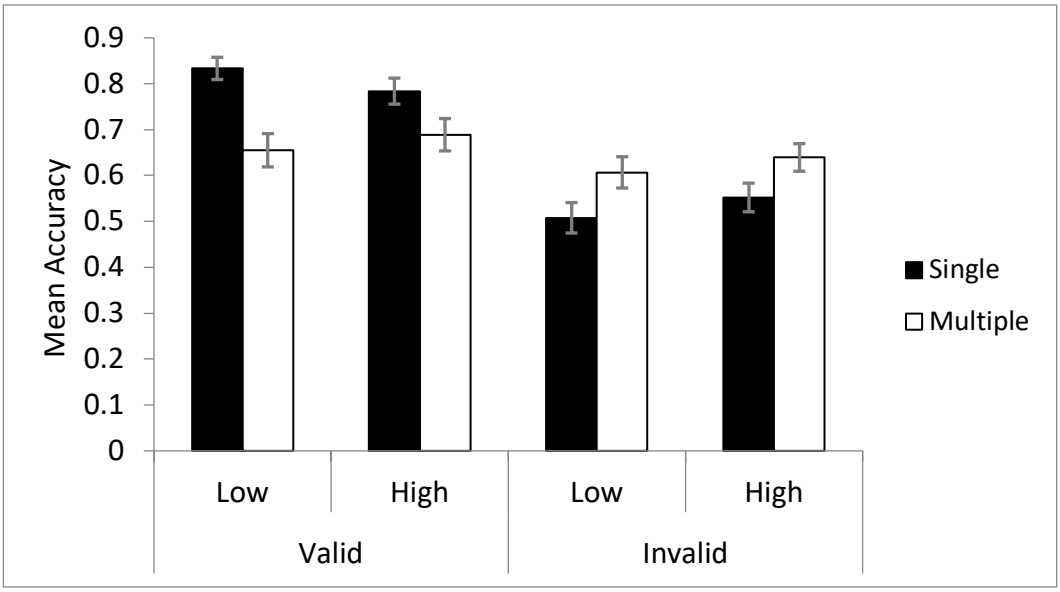


Figure 5.2. Mean accuracy as a function of model, anchor, and validity. Error bars represent standard errors.

Table 5.1. Mean Accuracy by model and validity

Model	**Validity**	**Mean**	**Std. Error**	**N**
Single	Valid	0.81	0.02	63
	Invalid	0.53	0.03	63
Multiple	Valid	0.67	0.03	63
	Invalid	0.62	0.03	63

5.2.4. Fluency Defined by Item RT

We again calculated a median composite RT for every participant, and divided their responses into two categories, fluent and disfluent trials. Consistent with the prior experiments, we found a significant effect of fluency on reanswer choices, $t(63) = -3.454$, $p = 0.001$.

Participants were more likely to select problems they solved less fluently to reanswer ($M = 0.35$, $sd = 0.31$) compared to those they solved fluently ($M = 0.28$, $sd = 0.33$).

5.2.5. Results Summary

For each of the four experiments, results with $p < 0.05$ are summarized in Tables 5.2 – 5.6. As can be seen from these tables, the results associated with FORs in Experiments 1 and 2 were similar. Furthermore, Experiments 2, 3, and 4 replicated most of the effects on composite RT, accuracy, and fluency from Experiment 1. Contrarily, the effect of number of models on reanswer choices only appeared in Experiment 1.

Table 5.2 A summary of results on FORs across four experiments. A check mark represents the presence of a significant effect. NA denotes the effect was not measured in the experiment.

FORs				
	E1	E2	E3	E4
Model	✔	✔	NA	NA
Anchor	✔	✔	NA	NA
Validity		✔	NA	NA
Model x Validity	✔	✔	NA	NA

Table 5.3 A summary of results on reanswer choices across four experiments.

Reanswer choices				
	E1	E2	E3	E4
Model	✔			

Table 5.4 A summary of results for composite RT across four experiments.

Composite RT				
	E1	E2	E3	E4
Model	✔	✔	✔	✔
Validity			✔	✔
Model x Validity	✔			

Table 5.5 A summary of accuracy results across four experiments.

Accuracy				
	E1	E2	E3	E4
Model		✔		
Validity	✔	✔	✔	✔
Model x Validity	✔	✔	✔	✔

Table 5.6 A summary of fluency-related results across four experiments.

Fluency				
	E1	E2	E3	E4
Fluency & FOR	✔	✔	NA	NA
FOR & reanswer	✔	✔	NA	NA
Fluency & reanswer	✔	✔	0.06	✔

To verify that including FOR did not change people's accuracy and composite RT, we compared the results from experiments with FOR judgements to those experiments without them, namely the control experiments. The first two experiments containing the FOR questions were treated as one group, whereas the control experiments were considered as the other group. This formed a 2 x (Anchor [low, high]) x 2 (Model [single, multiple]) x 2 (Validity [valid, invalid]) repeated measures ANOVA with group [FOR, control] as the between-subject factor. Due to some missing data on composite RT in Experiment 1, the FOR group comprised 105 participants instead of 128 participants. For composite RT, none of the independent variables interacted with group. Additionally, the effect of group on composite RT was non-significant, $F(1, 231) = 3.017$, $p = 0.084$, $\eta_p^2 = 0.013$. To examine reanswer choices, none of the independent variables interacted with group except for number of models. The interaction between model and group was significant, $F(1, 254) = 6.203$, $p = 0.013$, $\eta_p^2 = 0.024$. To decompose the interaction, the effect of model on reanswer choices was significant for the FOR group (mean difference = 0.38, $p = 0.001$), but not for the control group (mean difference = 0.003, $p = 0.786$). This significant effect within the FOR group was present because the effect was found in Experiment 1, but not in Experiment 2.

Nevertheless, the effect of group on reanswer choices was non-significant, $F(1, 254) = 2.642$, $p = 0.105$, $\eta_p^2 = 0.010$. For accuracy, there was no significant interaction between group and any of the independent variables. The effect of group was again non-significant on accuracy, $F(1, 254) = 0.511$, $p = 0.475$, $\eta_p^2 = 0.002$. We also compared the effect of fluency on reanswer choices between the FOR group and the control group. Again, there was no significant difference found between the two groups regarding the effect of fluency on reanswer choices, $F(1, 254) = 2.93$, $p = 0.88$, $\eta_p^2 = 0.011$. By comparing the dependent variables (composite RT, accuracy,

reanswer choices, and answer fluency) between experiments with FORs and without FORs, we can tentatively conclude that including FORs would not affect people's performance on the task.

5.3. Discussion

The comparison between experiments containing FORs and those without FORs showed that people's accuracy and composite RT were not affected when FOR judgements were included in the experiments. Moreover, there was no anchoring effect on reanswer choices across the four experiments (with and without FOR questions), which suggested incorporating the FOR question did not alter people's reanswer choices. We can conclude that including FOR judgements in experiments does not affect people's performance on the reasoning task, that is, their accuracy, composite RT, and possibly their reanswer choices.

The only significant difference found between the manipulated and the control experiments was the effect of number of models on reanswer choices. This difference was mainly driven by Experiment 1, because the effect of number of models on reanswer choices was only observed in Experiment 1, but not in the following experiments. Therefore, Experiments 1 and 2 overall produced a significant effect on reanswer choices by the number of models, but the effect on reanswer choices was non-significant in both Experiments 3 and 4, which resulted in a significant interaction effect.

We proposed in Hypothesis B that variables affecting fluency would influence FOR, which in turn would predict people's subsequent reanswer choices. The data from Experiment 1 supported this hypothesis, but the remaining three experiments failed to demonstrate the anticipated effects on reanswer choices. More specifically, the effect of number of models on reanswer choices was absent, although the relationship between fluency and reanswer choices held true. In other words, overall less fluent problems were more likely to be reanswered, but

people did not choose to reanswer multiple-model syllogisms over single-model counterparts, even though the former were solved slower.

To solve the conundrum, we categorized all of the syllogisms used in each experiment into one of four conditions: Single & Fluent, Single & Disfluent, Multiple & Fluent, and Multiple & Disfluent. Each syllogism was either a single-model or multiple-model problem with a composite RT either faster or slower than the median composite RT. The number of syllogisms in each condition are displayed in Table. 5.7. As shown by the table, the majority of the single-model syllogisms were fluent, and the multiple-model problems were less fluent. However, there still existed a substantial proportion of single-model problems that were disfluent, and multiple-model problems that were fluent. These data suggested that the disfluent single-model problems might have counteracted the fluent ones, mitigating the fluency effect produced by the models on reanswer choices. That is, the outcome turned out to be null for reanswer choices because single-model problems were not necessarily fluent; thus, its predictability on reanswer choices was reduced. Similar reasoning was applied to the multiple-model problems. On the other hand, the item-based RT analysis was to compare each individual's composite RT to their own median composite RT, which made half of the problems faster than the median (i.e., fluent) and the other half slower than the median (i.e. disfluent) regardless of which model was involved. Again, single-model syllogisms were not always fluent in comparison.

Table 5.7 Number of syllogisms in each condition for all four experiments.

	Single. Fluent	**Single. Disfluent**	**Multiple. Fluent**	**Multiple. Disfluent**
E1	476	421	435	477
E2	556	444	449	545
E3	568	422	426	556
E4	577	435	437	564

The participants in the present study were paid instead of receiving course credit as in the previous experiments. The probability of reattempting the problems was about 10% higher in the current experiment compared to Experiments 1 - 3, but none of the manipulated variables produced any significant effects on people's reanswer choices, replicating previous results. This finding contradicted our prior account postulating that the absence of the anchoring effect was due to a floor effect of the reanswer choices.

Chapter 6. General Discussion

A series of four experiments investigated the effect of FORs on reanswer choices by directly manipulating FORs. We attempted to answer the question: without the effect of answer fluency, is a manipulation of FOR sufficient to predict people's reanswer choices? To this end, we examined a cue that consistently affected answer fluency (i.e., number of models) and a cue that directly influenced FOR without affecting answer fluency (i.e., the size of anchors). We proposed two hypotheses, attempting to account for the relationships amongst answer fluency, FOR, and reanswer choice. Hypothesis A: cues can either directly or indirectly (through the effect of answer fluency) influence FORs, which in turn predict people's subsequent reanswer choices. According to this hypothesis, we predicted higher FORs for single-model syllogisms and for higher-anchor problems, and that people would be less likely to choose these problems to reanswer. The alternative, hypothesis B, suggested that only cues affecting answer fluency and FORs would subsequently predict people's reanswer choices. More specifically, we predicted both the size of anchors and number of models would influence FORs, but reanswer choices would only be affected by the latter, because it would also affect answer fluency.

6.1. Answer Fluency, FOR, and Reanswer Choices

Our data showed that people provided higher FORs for problems paired with high anchors and for single-model syllogisms, consistent with both of the hypotheses. In one of four experiments, people were less likely to choose single-model syllogisms to reanswer because they were more fluent than their multiple-model counterparts, whereas there was a lack of anchoring effect on reanswer choices. Therefore, the results from only Experiment 1 supported hypothesis B, indicating that FOR can predict people's subsequent reanswer choices only if it is affected by answer fluency, but not directly through other cues. However, this effect of models on reanswer

choices was absent in the remaining three experiments. To solve the conundrum, we counted the number of syllogisms and classified them into one of four categories: fluent single-model problems, disfluent single-model problems, fluent multiple-model problems, and disfluent multiple-model problems. We found that a substantial number of single-model items were actually less fluent than their multiple-model counterparts, which was not anticipated. Similarly, many multiple-models were fluent instead of being disfluent. If an individual reanswered the fluent problems, there would be a mix of single- and multiple-model items amongst their reanswer choices. In other words, even if one was more likely to reanswer fluent than disfluent problems, this might not translate into an effect of model on fluency. Therefore, reanswer choices are determined by the relative fluency of problems for an individual, according to his or her own median composite RT.

We also considered two other possible explanations to explain the lack of anchoring effects and the effect of number of models on reanswer choices in the last three experiments. One possible explanation is that the low likelihood of reanswering might have impeded us from observing the effect of anchoring or models on reanswer choices. The syllogisms used in the experiments might be too difficult for the participants to solve, resulting in a floor effect on reanswer choices. Nonetheless, this explanation was made less likely given that the probability of reanswering was increased to approximately 30% in Experiment 4, and neither the effect of number of models nor the size of anchors on reanswer choices were observed. The second account posits that FOR might not predict people's subsequent intention towards reattempting the problems as measured by their reanswer choices. This explanation was also not likely considering that FOR was associated with reanswer choices on the individual level according to the item-based RT analysis. The observation that people were more likely to reanswer

problems that were given lower FORs based on the item-based RT analysis suggested FOR was, in fact, able to predict reanswer choices. Thus, this possible explanation was not supported by our results.

6.2. FOR and Accuracy

In support of our hypotheses on FORs, variables such as the size of anchors and number of models consistently predicted FORs regardless of accuracy in these experiments. That is, high anchor values and single-model syllogisms elevated people's FOR judgements, but their accuracy on the reasoning task was unchanged. Our results were similar to previous findings that confidence judgements and accuracy are dissociated in the realm of reasoning (Bajšanski et al., 2014; Quayle & Ball, 2000; Shynkaruk & Thompson, 2006; Thompson et al., 2011).

6.3. The Outcome of Including FOR

We further tested the outcome of including FOR judgements in the experiments in order to rule out the concern that the act of adding FOR may change people's performance on the reasoning task. By comparing the results from experiments containing the FOR question to those without it, we showed that people behaved similarly with and without FOR judgements regarding accuracy and composite RT. These results suggest that including the FOR question does not inflate reflective behaviours such as longer RTs and increased accuracy. The results associated with reanswer choices were less clear due to inconsistent results across the four experiments. Nevertheless, the absence of the anchoring effect on reanswer choices across four experiments suggested incorporating FOR judgements in the experiments did not alter people's reanswer choices.

6.4. Comparison to Metamemory Literature

Our experiments partially replicated the research conducted by Yang et al. (2018). Both studies showed that people's subsequent judgements were susceptible to the anchoring effect, even though different measures of judgements were employed in the two research paradigms. More specifically, we assessed people's FORs, and Yang and colleagues measured JOLs. We were unable to find the effect of anchoring on reanswer choices as Yang et al. did for restudy choices, possibly for the reasons discussed above. In addition, we attempted to adhere to the procedure of the previous research as closely as possible, but we used a different task and measures for the purpose of our experiments. The task involved in our experiments was solving syllogisms, whereas Yang et al.'s task was memorizing word-pairs. The former requires participants to think about the problems, which was more cognitively demanding than memorization. The anchor value might act as a superficial cue that only affected people's judgement represented by the measure, namely FOR, but not their cognitive process or experience related to problem-solving.

6.5. Limitations and Future Research

The experiments have a few caveats that need to be addressed. We asked the participants to provide a reanswer choice for each problem, and the question was phrased as the following: "To improve your overall score, would you like to attempt to solve this problem again after you have solved all of the problems?" The phrasing of this question might measure people's motivation more than their actual behaviours. We adopted the methodology of asking for participants' reanswer choices instead of making them solve the problems again in order to replicate prior research (Yang et al., 2018). The drawback is that this methodology prevented us from comparing our data to previous meta-reasoning work that measured participants' actual

reanswering behaviours (Thompson et al., 2011; Thompson et al., 2013; Thompson & Johnson, 2014). In these prior studies, participants solved each problem twice, and their rethinking time and answer change were denoted as reanswering behaviours. In contrast, our experiments measured people's reanswer choices, which might be different than rethinking and answer-change behaviours. Thus, we need to take into account the possibility that the relationship between FOR and reanswering choices is not completely parallel with the relationship between FOR and reanswer behaviours.

Additionally, participants were not timed under a deadline for their intuitive responses. Prior research indicates that FORs increase over time (Shynkaruk & Thompson, 2006; Thompson et al., 2011). In our experiments, participants might have reflected on the problems and provided higher FORs, which in turn led to lower probability of reanswer choices. We tried to control for any deliberate responses by eliminating trials that participants failed to answer intuitively according to their self-reports, but the validity of the self-reports remained uncertain. Moreover, people's likelihood of reanswering was relatively low in the first three experiments, which made the interpretation of the effect on reanswer choices more complicated. To address these limitations, future research could incentivize the participants to be more engaged in the task by rewarding them for each question they answer correctly. This rewarding method could motivate participants to reanswer more problems. Another direction would be measuring people's actual reanswering behaviours with the same paradigm and compare the results with previous research.

Chapter 7. Conclusion

The data from a series of four experiments supported the hypothesis that only cues affecting both answer fluency and FOR can predict people's subsequent reanswer choices to some extent. We could not completely verify the above hypothesis, however, because the cue that affected answer fluency (i.e., number of models) in our experiments was not as reliable when examining reanswer choices. Future research will need to use a more effective cue to investigate the relationships amongst answer fluency, FOR, and reanswer choices. We further provided evidence supporting previous research that confidence judgements are not related to accuracy of the reasoning task (Bajšanski et al., 2014; Quayle & Ball, 2000; Shynkaruk & Thompson, 2006; Thompson et al., 2011). Additionally, we showed that adding FOR judgements to the experiments did not change people's performance on the reasoning task, addressing the concern about the use of FORs in experiments.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2017.05.004

Bajšanski, I., Močibob, M., & Valerjev, P. (2014). Metacognitive judgments and syllogistic reasoning. *Psychological Topics*, *23*(1), 143-166.

Bajšanski, I., Žauhar, V., & Valerjev, P. (2018). Confidence judgments in syllogistic reasoning: the role of consistency and response cardinality. *Thinking & Reasoning*, *25*(1), 1-34. https://doi.org/10.1080/13546783.2018.1464506

Ball, L. J., Phillips, P., Wade, C. N., & Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: eye-movement evidence for selective processing models. *Experimental Psychology, 53*(1), 77–86. http://doi.org/10.1027/1618-3169.53.1.77

Bara, B., Bucciarelli, M., & Johnson-Laird, P. (1995). Development of syllogistic reasoning. *The American Journal of Psychology,108*(2), 157-193. http://doi.org/10.2307/1423127

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610-632. https://doi.org/10.1016/0749-596X(89)90016-8

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55-68. http://doi.org/10.1037/0096-3445.127.1.55

Chapman, G. B., & Johnson, E. J. (1999). Anchoring, activation, and the construction of values. *Organizational Behavior and Human Decision Processes*, *79*(2), 115–153. https://doi.org/10.1006/OBHD.1999.2841

Copeland, D., & Radvansky, G. (2004). Working memory and syllogistic reasoning. *The Quarterly Journal of Experimental Psychology Section A*, *57*(8), 1437-1457. https://doi.org/10.1080/02724980343000846

England, B. D., & Serra, M. J. (2012). The contributions of anchoring and past-test performance to the underconfidence-with-practice effect. *Psychonomic Bulletin & Review, 19*(4*)*, 715–722. http://doi.org/10.3758/s13423-012-0237-7

Evans, J. St. B. T., Handley, S. J., Harper, C. N. J., & Johnson-Laird, P. N. (1999). Reasoning about necessity and possibility: a test of the mental model theory of deduction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1495–1513. http://doi.org/10.1037/0278-7393.25.6.1495

Frederick, S. W., & Mochon, D. (2012). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General*, *141*(1), 124–133. https://doi.org/10.1037/a0024006

Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35-42. https://doi.org/10.1016/j.socec.2010.10.008

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 22. http://doi.org/10.1037/0278-7393.29.1.22

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition, 16*(1), 1–61.https://doi.org/10.1016/0010-0277(84)90035-0

Johnson-Laird, P. N. and Byrne, R. M. J. (1991). *Deduction.* Hillsdale, N. J.: Lawrence Erlbaum.

Johnson-Laird, P. N., & Steedman, M. (1978). The psychology of syllogisms. *Cognitive psychology*, *10*(1), 64-99. https://doi.org/10.1016/0010-0285(78)90019-1

Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107*(4), 852-884. http://doi.org/10.1037/0033-295X.107.4.852

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), The Cambridge handbook of consciousness (pp.289-325). New York, NY: Cambridge University Press.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174-179. https://doi.org/10.3758/PBR.15.1.174

Mussweiler, T., & Strack, F. (2000). The use of category and exemplar knowledge in the solution of anchoring tasks. *Journal of Personality and Social Psychology*, *78*(6), 1038–1052. https://doi.org/10.1037/0022-3514.78.6.1038

Nelson, T. O. (1990). Metamemory: a theoretical framework and new findings. In G.H. Bower (Ed). *Psychology of Learning and Motivation*(Vol. 26, pp. 125-173). New York, NY: Academic Press.

Prowse Turner, J. A., & Thompson, V. A. (2009). The role of training, alternative models, and logical necessity in determining confidence in syllogistic reasoning. *Thinking & Reasoning*, *15*(1), 69-100. https://doi.org/10.1080/13546780802619248

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). Retrieved from http://www.pstnet.com

Quayle, J. D., & Ball, L. J. (2000). Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. *Quarterly Journal of Experimental Psychology, 53*(4), 1202–1223. https://doi.org/10.1080/713755945

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615. https://doi.org/10.3758/PBR.16.3.550

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory and Cognition, 34*(3), 619–632. https://doi.org/10.3758/BF0319358

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204-221. http://doi.org/10.1037/0278-7393.26.1.204

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*(3), 437–446. https://doi.org/10.1037/0022-3514.73.3.437

Thompson, V. A., Evans, J. S. B. T., & Campbell, J. I. D. (2013). Matching bias on the selection task: it's fast and feels good. *Thinking and Reasoning*, *19*(3-4), 431-452. https://doi.org/10.1080/13546783.2013.820220

Thompson, V. A., Turner, J. A. P., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*(3), 107-140. https://doi.org/10.1016/j.cogpsych.2011.06.001

Thompson, V. A., Turner, J. A. P., 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*(2), 237-251. https://doi.org/10.1016/j.cognition.2012.09.012

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10. 1017/CBO9780511809477.002

Undorf, M., & Erdfelder, E. (2011). Judgments of learning reflect encoding fluency: conclusive evidence for the ease-of-processing hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1264-1269. http://doi.org/10.1037/a0023719

Wansink, B., & Sobal, J. (2007). Mindless eating: The 200 daily food decisions we overlook. *Environment and Behavior*, *39*(1), 106-123. https://doi.org/10.1177/0013916506295573

Wetherick, N. E., & Gilhooly, K. J. (1995). "Atmosphere," matching, and logic in syllogistic reasoning. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues, 14*(3), 169-178. http://doi.org/10.1007/BF02686906

Yang, C., Sun, B., & Shanks, D. R. (2018). The anchoring effect in metamemory monitoring. *Memory & Cognition*, *46*(3), 384-397. https://doi.org/10.3758/s13421-017-0772-6

Zhao, Q. (2012). Effects of accuracy motivation and anchoring on metacomprehension judgment and accuracy. *Journal of General Psychology, 139*(3), 155–174. https://doi.org/10.1080/00221309.2012.680523

Zhao, Q., & Linderholm, T. (2011). Anchoring effects on prospective and retrospective metacomprehension judgments as a function of peer performance information. *Metacognition and Learning, 6*(1), 25–43. https://doi.org/10.1007/s11409-010-9065-1

Appendix

In a syllogism, a quantifier indicates the scope of the given sets. For example, words like "all" and "no" are categorized as universal, whereas "some" is referred to as particular. The quantifier of each statement may be affirmative or negative; that is, the quantifier can either affirm that one group belongs to another group or negates it. Therefore, the quantifiers can have four possible forms, which are also known as moods (Johnson-Laird & Bara,1984). Examples with the common single-letter abbreviations are listed as follows:

All philosophers are logicians — affirmative-universal (A)

Some teachers are painters— affirmative-particular (I)

No musicians are engineers — negative-universal (E)

Some gardeners are not models — negative-particular (O)

Another factor in a syllogism that needs to be considered is "figure". Figure refers to the sequence in which the A, B and C terms are presented. There are four types of figure, which are listed below:

A-B B-A A-B B-A

B-C C-B C-B B-C

Both mood and figure affect performance (Johnson-Laird & Steedman, 1978). Specifically, the difficulty of the syllogisms is dependent on mood and figure interacting with each other.