Psychological Topics, 28 (2019), 1, 37-72
Original Scientific Paper UDC: 159.955.072
doi:https://doi.org/10.31820/pt.28.1.3

Fluency and Feeling of Rightness: The Effect of Anchoring and Models

Selina Wang and Valerie Thompson
University of Saskatchewan, Saskatoon, Canada

Abstract

Feeling of Rightness (FOR) is a metacognitive experience accompanying people's intuitive answers that predicts the probability of subsequently changing answers (Thompson, Prowse Turner, & Pennycook, 2011). Previous research suggested FOR judgments are influenced by cues such as fluency, i.e., the ease with which an answer comes to mind. In the current paper, we examine the relationship between FOR, fluency, and answer changes; in particular, we were interested in whether answer fluency drives the effect of FOR on subsequent behaviours pertaining to answer changes. Reasoners ($N = 64$) in each of four experiments were asked to determine the validity of 32 syllogisms that consisted of single-model and multiple-model problems. In addition, each problem was randomly paired with a question containing either a high anchor value (80% or 90%) or a low anchor value (10% or 20%). In the first two experiments, reasoners then provided a FOR rating on a scale from 0 to 100 and indicated whether they would like to attempt to re-answer the question. The last two experiments served as the control experiments in which the FOR judgements were removed. The anchoring manipulation affected FOR judgments but not re-answer choices; it also did not affect answer fluency. Thus influencing FOR without affecting answer fluency had no effect on people's subsequent re-answer choices. In contrast, fluency was a reliable predictor of both FOR and re-answer choices. That is, when answers came to mind slowly, FORs were lower and people were more likely to choose to re-answer the problems. Thus, fluency appears to mediate the relationship between FOR and re-answer choices.

Keywords: feeling of rightness, fluency, meta-reasoning, syllogistic reasoning, anchoring, mental models

✉ Valerie Thompson, Department of Psychology, University of Saskatchewan, 9 Campus Drive, Saskatoon S7N 5A5, SK, Canada. E-mail: *valerie.thompson@usask.ca*

This research was supported by the Natural Science and Engineering Research Council of Canada.

Introduction

According to researchers at Cornell University, we make 226.7 decisions about food alone every day (Wansink & Sobal, 2007). Given the substantial number of 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 & Narens, 1990). Its function is analogous to a working thermostat that passively measures room temperature and controls the initiation and termination of the furnace (Ackerman & Thompson, 2017). This paper addresses issues of metacognitive monitoring and control as they apply to reasoning, which is also denoted as meta-reasoning.

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. Their analysis was initially based on the metamemory literature. Monitoring judgments, such as the Judgement of Learning (JOL), measure peoples' estimates of how well they have learned a particular piece of information, which, in turn, directly influences their subsequent study choices (Metcalfe & Finn, 2008; Son & Metcalfe, 2000). Specifically, when asked to remember pieces of information, people are less likely to restudy the ones that they believed they are likely to be able to recall on a later test. Analogously, 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, Prowse Turner, & Pennycook, 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.

Relationship between FOR and Re-Answer 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 change their initial answer 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, 2013; Thompson & Johnson, 2014). In other words,

if reasoners were confident that their answers were correct (i.e., high FOR), they spent less time rethinking the problems and were less likely to change their original answers.

Predictors of FOR

Metacognitive judgements such as Judgements of Learning (JOLs) are thought to be based not 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 correlate with accuracy. 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 (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 reasoners' 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, which is the ease with which an answer comes to mind. Most of the current methods that have been used to identify the cues that underlie FOR are correlational: Faster responses are generally associated with higher FORs (e.g., Thompson et al., 2011). There are also experimental ways to identify such cues. These methods involve manipulating variables that will speed or slow initial responses, and then show that the "down-stream" measures of rethinking times and answer-changes change in tandem with the manipulation. That is, as described below, prior attempts to manipulate FOR were indirect in that they also affected answer fluency (Thompson et al., 2011, 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; Experiment 3) manipulated FOR using base-rate 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). As predicted, there were longer rethinking times and a greater probability of answer change for conflict than non-conflict problems. In addition, 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; Experiment 4; 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 deciding whether to turn over each card. Reasoners 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). Thus, 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. However, because FOR was manipulated by manipulating fluency, it was difficult to tease apart the effects generated by FOR from answer fluency. Thus, it is not clear which of the two variables is the proximal cause of rethinking times and answer change behaviours.

To summarize, previous research has demonstrated that FOR predicts people's re-answering behaviours (Thompson et al., 2011, 2013); both correlational and experimental data showed that variables that affected FOR also affected the downstream behavioural effects (i.e., answer change). Therefore, the question that remains to be answered is whether manipulating FOR per se would be sufficient to predict people's subsequent re-answering behaviours independently of answer fluency. Therefore, the goal of the following experiments was to examine the role of FOR in relation to subsequent re-answering choices by using a cue that directly manipulated FOR without affecting fluency, and a cue that indirectly affected FOR (via the effect of answer fluency). To this end, as is explained below, we employed an "anchoring" manipulation as a cue that could directly influence FOR without affecting fluency.

Anchoring

Anchoring occurs when people incorporate a previously encountered value into a subsequent estimate, even when that value is irrelevant to the estimate. For example, people 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; Mussweiler & Strack, 2000; Strack & Mussweiler, 1997). 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 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 to be darker after walking out of 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). Although, it has not been widely investigated as a potential cue to examine metacognitive judgments, the available data suggest that anchoring can be used to manipulate metacognitive judgements such as JOL (England & Serra, 2012; Yang, Sun, & Shanks, 2018; Zhao, 2012; Zhao & 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 in which the anchor values were said to represent peer, 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, that is, they were uninformative, the relationship between anchoring and the metacognitive judgements was less clear (England & Serra, 2012; Zhao, 2012; Zhao & Linderholm, 2011). To address this gap in the research and to elucidate the role of anchoring in metamemory monitoring and control, Yang et al. (2018) conducted a series 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. Participants then provided a JOL score from 0 to 100, indicating the probability that they would be able to remember the pair in 5 minutes. 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 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 (2018) 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' behaviour.

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. In addition, as described below, we also manipulated a variable that we expected would affect fluency.

Models in Syllogistic Reasoning

The materials for this study consisted of three-term syllogisms. 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. To evaluate its validity (i.e., whether it follow logically from the premises), people test if the conclusion is consistent with the model. To be thorough, they should then test the conclusion against possible alternative models of the premises; if the conclusion is not consistent with all of these, it should be rejected. In practice, however, people often neglect this step and end up accepting invalid conclusions. A valid conclusion is one that is consistent with all possible models of the premises; 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 impact the difficulty of the problem. We manipulated this variable in the current experiments, because we assumed that 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 studies found that people spent more time solving multiple-model syllogisms than their single-model counterparts (e.g., Copeland & Radvansky, 2004), suggesting that the latter may be more fluent. Given that response time is 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 potentially examine the effect of FOR on re-answer 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 did find that single-model syllogisms were given higher on confidence judgments than their multiple-model counterparts. Therefore, they 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 et al., 2014; Quayle & Ball, 2000; Shynkaruk & Thompson, 2006; Thompson et al., 2011). Therefore, we predicted that FORs would be higher for single-model and fluent problems than their counterparts, regardless of differences in accuracy.

Summary

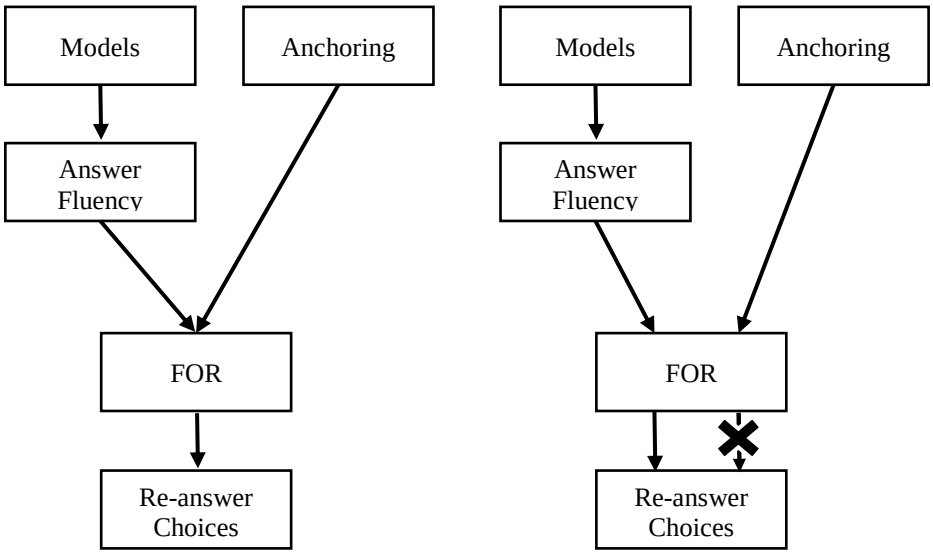
We attempted to investigate the role of FOR in predicting re-answer choices in syllogistic reasoning. To this end, we exploited a cue that was predicted to directly affect FOR, which was the size of anchors, and a cue that was predicted to influence FOR through the effect of answer fluency, which was 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 re-answer 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 re-answer choices differed from previous studies, which examined people's actual behaviours of reconsideration and answer change (Thompson et al., 2011, 2013), but was similar to the measures used in Yang et al.'s anchoring study (2018). 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 re-answer choices (see Figure 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 re-answer choices as illustrated in Figure 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 re-answer these problems. Hypothesis B: only cues that affect the experience of answer fluency (i.e., number of models) would predict subsequent re-answer choices as depicted in Figure 1b. Those cues that do not influence answer fluency may affect FOR, but they would not have any effects on re-answer choices. Therefore, answer fluency is the key factor for predicting re-answer choices. According to this hypothesis, we predicted that the number of models would influence answer fluency and FOR, which in turn would affect re-answer choices. On the other hand, anchoring would only affect FOR, but not re-answer 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. Flowcharts depicting two hypothesized paths of FORs in the current experiments.

Experiment 1

The paradigm used in Experiment 1 closely matched the study conducted by Yang et al. (2018). We examined the effect of uninformative anchors on FORs and re-answer choices. Reasoners were instructed to solve the syllogisms intuitively in order to be consistent with previous work on FOR (Thompson et al., 2011, 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), re-answer choices, and accuracy. To obtain a proxy measure of answer fluency, we summed response time and reading time.

Method

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.

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, which were presented using the E-Prime 2.0 Software Tools program (Psychology Software Tools, Pittsburgh, PA, 2012). 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 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.

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 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

[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.

the occupation-related content in each premise pair was accompanied by two valid and two invalid conclusions.

Table 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 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.

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. On each trial, participants saw two 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 re-answer 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.

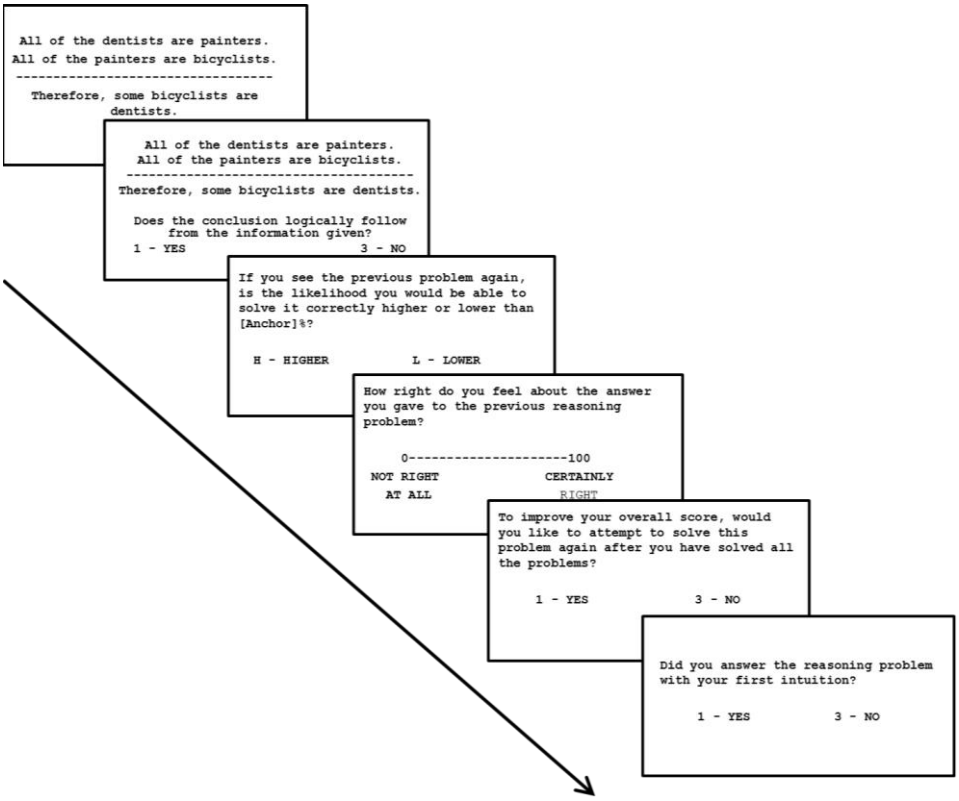


Figure 2. The trial progression for Experiment 1.

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, re-answer-choices, composite RT (the sum of response time and reading time), and accuracy. Because the focus of our study was on the relationship between FOR, fluency, and re-answer choices, the accuracy analyses are reported in Appendix B. Results with $p < .05$ were reported as significant. Paired *t*-tests were used to examine the simple main effects of the interactions.

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 3. Consistent with our hypotheses, syllogisms paired with low anchors were rated lower on FOR ($M = 81.61$, $SD = 1.41$) than high-anchor syllogisms ($M = 83.81$, $SD = 1.32$), $F(1,63) = 7.559$, $p = .008$, $\eta_p^2 = .107$. As predicted, FORs were higher for single-model syllogisms ($M = 84.92$, $SD = 1.28$) than for multiple-model syllogisms ($M = 80.51$, $SD = 1.45$), $F(1,63) = 30.687$, $p < .001$, $\eta_p^2 = .328$. There was also a significant interaction between model and validity, $F(1,63) = 7.721$, $p = .007$, $\eta_p^2 = .109$. Values for the interaction are presented in Table 2. People gave higher FORs for single-model than for multiple-model syllogisms when the problems were valid (+5.88; $t(63) = 5.019$, $p < .001$), but this difference was smaller when the problems were invalid (+3.00; $t(63) = 4.290$, $p < .001$).

[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.

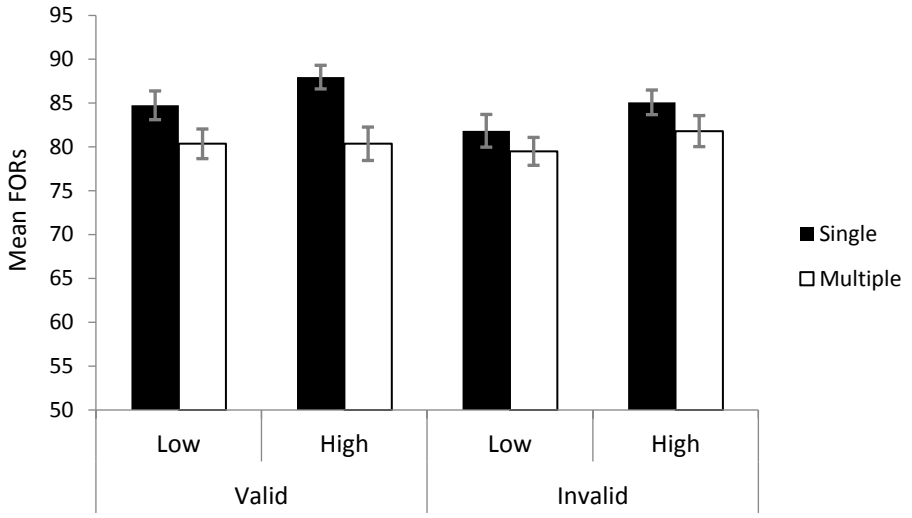


Figure 3. Mean FORs in Experiment 1 as a function of anchor, model and validity. Error bars represent standard errors.

Table 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

Re-Answer Choices. The overall mean probability of re-answering was 0.21. The data are plotted in Figure 4. Consistent with Hypothesis A and B, participants were less likely to re-answer single-model syllogisms (M = 0.18, SD = 0.03) than multiple-model syllogisms (M = 0.23, SD = 0.04), $F(1,63) = 8.845$, $p = .004$, $\eta_p^2 = .123$. The interaction between model and validity was also significant, $F(1,63) = 5.010$, $p = .029$, $\eta_p^2 = .074$. Values for the interaction are presented in Table 3. When the problems were invalid, people were more likely to re-answer multiple-model syllogisms than single-model ones (+0.090; $t(63) = 4.797$, $p < .001$), but this difference was not found in valid problems (+0.015; $t(63) = 0.581$, $p = .564$). Contrary to Hypothesis A, the main effect of anchor was not significant, $F(1,63) = 0.007$, $p = .933$, $\eta_p^2 < .001$.

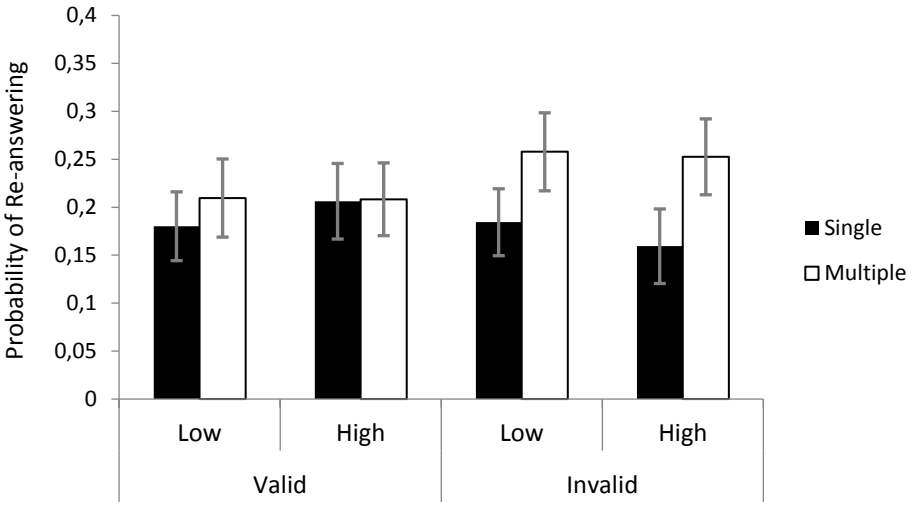


Figure 4. Probability of re-answering in Experiment 1 as a function of anchor, model and validity. Error bars represent standard errors.

Table 3

Probability of Re-Answering 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

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), $F(1,40) = 13.955$, $p = .001$, $\eta_p^2 = .259$. Furthermore, the model and validity interaction was marginally significant, $F(1,40) = 3.833$, $p = .057$, $\eta_p^2 = .087$. Values for the interaction are displayed in Table 4. For valid syllogisms, participants solved single-model problems faster than multiple-model ones (-2.93; $t(40) = -3.380$, $p = .002$), whereas the difference between composite RT was marginally significant for invalid syllogisms (-1.27; $t(40) = -1.863$, $p = .07$). The size of the anchors had no effect on composite RT, $F(1, 40) = 1.266$, $p = .267$, $\eta_p^2 = .031$.

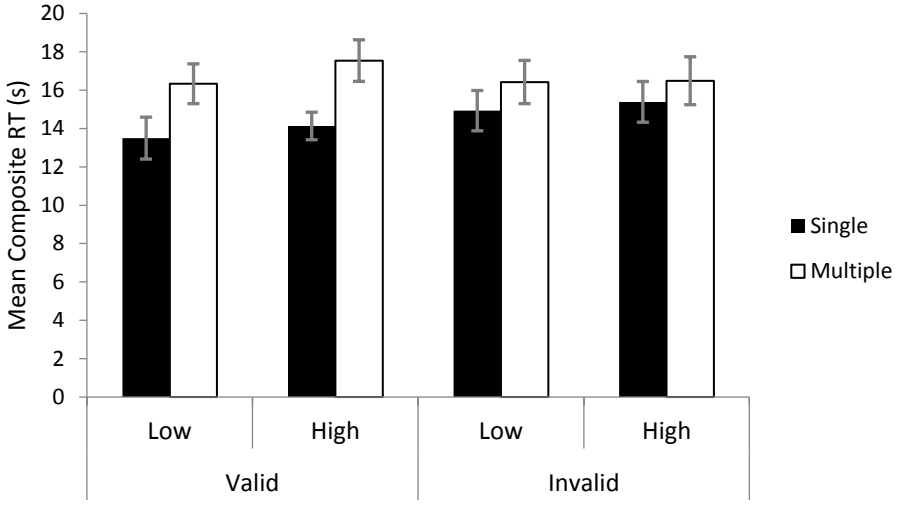


Figure 5. Mean composite RT for Experiment 1 as a function of anchor, model and validity. Error bars represent standard errors.

Table 4

Mean Composite RT 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

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]. 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, 2013), fluently produced answers were given higher FORs (M = 83.54, SD = 10.58) than their disfluent counterparts (M = 81.11, SD = 11.13),

[3] We also computed a median reading RT for each of the 41 participants whose reading time were 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), $t(40) = 5.963$, $p < .001$. The effect of fluency on re-answer choices was only marginally significant, $t(40) = -1.748$, $p = .088$. However, the trend was similar to the composite RT data, 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).

$t(40) = 4.937$, $p < .001$. 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 ($M = 83.58$, $SD = 10.92$), $t(46) = -3.083$, $p = .003$. We also examined the fluency effect on re-answer choices, which was also significant, $t(40) = -1.998$, $p = .05$. 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.

Discussion

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 re-answer. These results were consistent with both of our hypotheses, because both hypotheses support that FOR can predict re-answer choices when it is influenced by answer fluency. Evidence from item-based RT analysis also demonstrated the expected relationship between FORs and re-answer choices in that FORs were lower for the problems people were willing to reattempt.

Size of Anchors. We were able to demonstrate the anchoring effect on FOR (Figure 5) 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 re-answer 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 re-answer 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 & Linderholm, 2011). It is plausible that answer fluency is the source of the actual experience, which in turn predicts peoples' subsequent re-answer choices. On the other hand, cues that directly affect FOR may only influence the judgement of the FOR, but not affect the experience associated with solving the problem. We postulate that this latter experience drives re-answer choices. Our item-based RT analysis was consistent with this hypothesis: 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 re-answer choices. Thus, affecting FOR without affecting fluency may remove its behavioural consequences.

An alternative explanation is that the anchoring effect on re-answer 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 re-answer in the current experiment was only about 20% compared to the likelihood of restudying in previous research which was about 36%. People's re-answer 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.

In the next experiment, the aim was to confirm the effects of number of models and size of anchors on re-answer 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 re-answer.

Experiment 2

The goal of Experiment 2 was to replicate the relationships amongst cues, FORs and re-answer choices as found in the previous experiment while attempting to increase the number of re-answer choices.

Method

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.

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.

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, re-answer choices, composite RT (the sum of reading time and response time), and accuracy. Results with $p < .05$ were reported as significant. Paired *t*-tests were employed to reveal the simple main effects for significant interactions. The accuracy data are analyzed in Appendix C.

FOR. The mean FOR rating collapsing all levels was 83.82. The FOR data are plotted in Figure 6. 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), $F(1,63) = 8.238$, $p = .006$, $\eta_p^2 = .116$. The main effect of validity was significant, $F(1,63) = 7.623$, $p = .008$, $\eta_p^2 = .108$. 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, $F(1,63) = 6.663$, $p = .012$, $\eta_p^2 = 0.096$. The values for the interaction are displayed in Table 5. To decompose this interaction, people gave higher FORs to single-model syllogisms than to multiple-model ones when the problems were valid (+6.13; $t(63) = 6.392$, $p < .001$), but the difference was smaller for invalid problems (+2.24; $t(63) = 2.187$, $p = .032$), replicating Experiment 1.

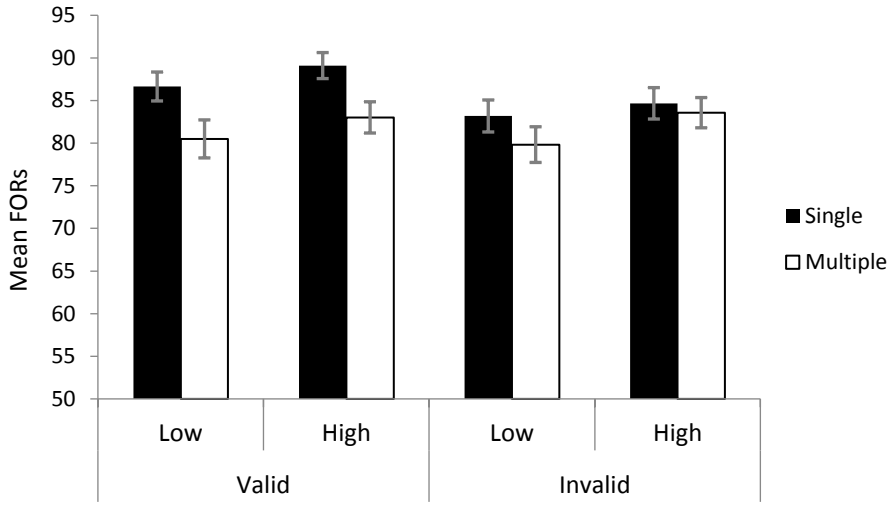


Figure 6. Mean FORs in Experiment 2 as a function of model, anchor, and validity. Error bars represent standard errors.

Table 5

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

Re-Answer Choices. The overall mean probability of re-answering was 0.20, only 0.01 lower than in Experiment 1. The data are illustrated in Figure 7. The values for the interaction are displayed in Table 6. Contrary to Experiment 1, there was a marginally significant interaction between model and anchor, $F(1,63) = 3.805$, $p = .056$, $\eta_p^2 = .057$. Participants were more likely to reattempt the multiple-model problems than their single-model counterparts for the low-anchor problems (+0.05; $t(63) = 2.379$, $p = .020$), but the pattern disappeared for the high-anchor problems (+0.01, $t(63) = 0.267$, $p = .790$). Again, there was no anchoring effect on re-answer choices, $F(1,63) = 0.244$, $p = .623$, $\eta_p^2 = .004$.

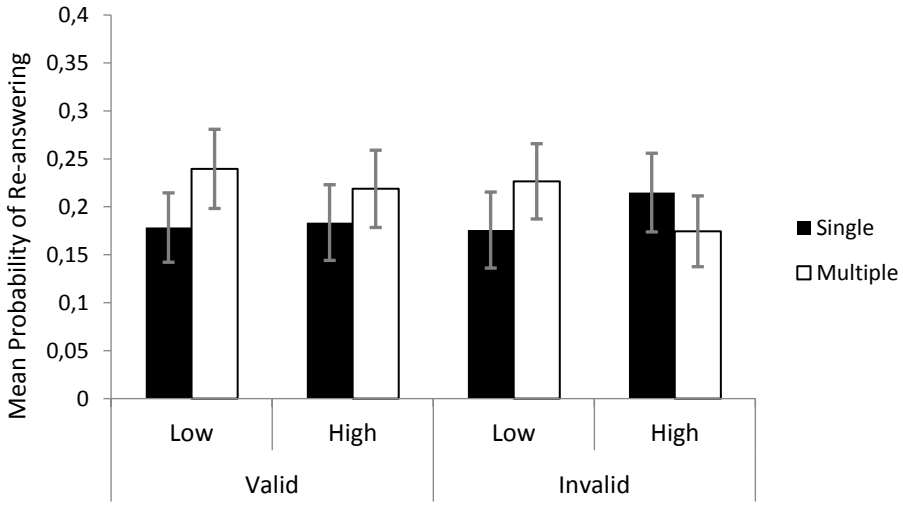


Figure 7. Mean probability of re-answering in Experiment 2 as a function of model, anchor, and validity. Error bars represent standard errors.

Table 6

Mean FORs by Model and Anchor

Model	Validity	Mean	Std. Error	*N*
Single	Valid	0.18	0.04	63
	Invalid	0.20	0.04	63
Multiple	Valid	0.23	0.04	63
	Invalid	0.20	0.04	63

Composite RT. Again, we computed the composite RT, which is the combination of participants' reading time and response time to the problems as our proxy for fluency. The overall mean composite RT was 20.02 seconds. The data are plotted in Figure 8. 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), $F(1,63) = 24.920$, $p < .001$, $\eta_p^2 = .283$. Again, the main effect of anchor on RTs was non-significant, $F(1,63) = 0.376$, $p = .542$, $\eta_p^2 = .093$.

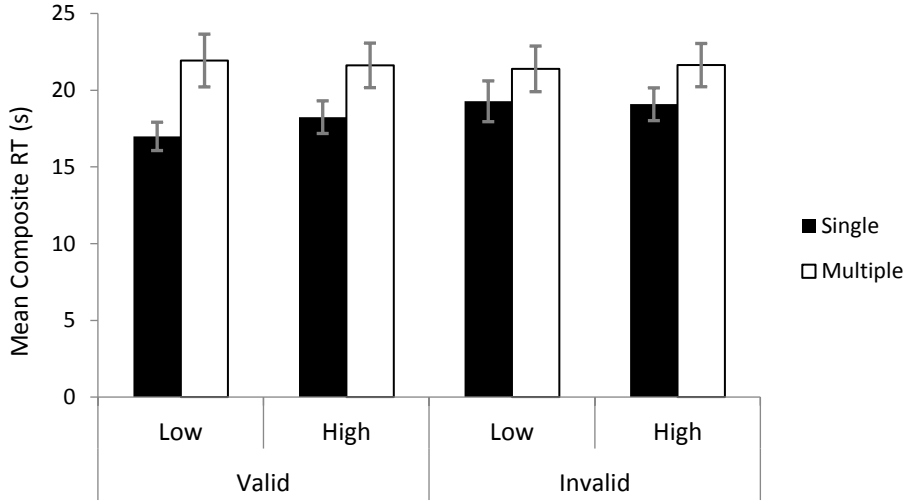


Figure 8. Mean composite RT in Experiment 2 as a function of model, anchor, and validity. Error bars represent standard errors.

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), $t(63) = 3.560$, $p = .001$. We again compared the FORs of participants' responses according to their re-answer choices. The items participants

preferred to re-answer were given lower FORs (M = 74.46, SD = 21.26) than those they were unwilling to reattempt (M = 84.57, SD = 13.54), $t(52) = -4.110$, $p < .001$. Additionally, there was a fluency effect on re-answer choices, $t(63) = -4.159$, $p < .001$. Participants were more willing to choose to re-answer disfluent problems (M = 0.24, SD = 0.29) than their fluent counterparts (M = 0.16, SD = 0.26).

Discussion

We attempted to remove the floor effect of re-answer 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 re-answering probabilities.

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, re-answer 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 re-answer choices unless they also affect answer fluency.

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. Unlike in Experiment 1, however, the model variable did not systematically affect re-answer choices.

In the current experiment, one of the manipulated variables, number of models, failed to produce any effects on re-answer choices, despite affecting both FOR and fluency. These data are difficult to explain. They also appeared to contradict the results from the item-based RT analysis. To foreshadow, we observed this null effect of number of models on re-answer choices in the next two experiments as well. We return to this issue in the General Discussion.

Experiments 3 & 4: Summary

The next two experiments served as control conditions, in which the FOR condition was omitted. This was done to address the concern that asking participants to make 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 re-answering. To avoid redundancy, and to simplify the presentation of the findings, we will focus on the few findings that differ from those reported above, and will report the results of the two experiments together.

Method

Participants. For Experiment 3, 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. For Experiment 4, 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.

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.

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: re-answer choices, composite RT, and accuracy. The accuracy data are reported in Appendices D and E.

Re-Answer Choices. The overall mean probability of re-answering was 0.20 in Experiment 3 and .32 in Experiment 4. Consistent with Experiment 2, the main effect of model on re-answer choices was non-significant, $F(1,63) = 0.169$, $p = .683$, $\eta_p^2 = .003$ and $F(1,63) = 0.002$, $p = .969$, $\eta_p^2 < .001$ respectively, as was the effect of anchor, $F(1,63) = 2.522$, $p = .117$, $\eta_p^2 = .038$ and $F(1,63) = 2.622$, $p = .110$, $\eta_p^2 = .040$.

Composite RT. Consistent with the previous experiments, we found a main effect of model on composite RTs, $F(1,63) = 44.777$, $p < .001$, $\eta_p^2 = .415$, and $F(1,63) = 38.140$, $p < .001$, $\eta_p^2 = .377$, such that participants responded to the single-model syllogisms more quickly than their multiple-model counterparts. Again, the main effect of anchor was non-significant, $F(1,63) = 0.008$, $p = .930$, $\eta_p^2 < .001$ and $F(1,63) = 2.622$, $p = .110$, $\eta_p^2 = .040$. The only finding that differed from the earlier ones was that instead of an interaction between validity and models, we observed a main effect of validity, $F(1,63) = 6.771$, $p = .012$, $\eta_p^2 = .097$ and $F(1,63) = 6.001$, $p = .017$, $\eta_p^2 = .087$ in both Experiment 3 and 4. The valid syllogisms were answered faster (M = 15.27, SD = 0.67 and M = 16.94, SD = 0.84) than the invalid ones (M = 16.22, SD =

0.79 and *M* = 18.10, *SD* = 0.97).

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* = .063, for Experiment 3 and *M* = 0.35, *SD* = 0.31 and *M* = 0.28, *SD* = 0.33, *t*(63) = -3.454, *p* = .001 in Experiment 4.

Conclusion. The results from the current experiments closely replicated Experiments 1 and 2, which suggest that including the FOR judgment did not change people's behaviour. Crucially, we observed the relationship between fluency and re-answer choices, even when the FOR judgment was excluded. Thus, the relationship between fluency and re-answer choices did not occur because the FOR question somehow caused people to engage in analytic thinking.

Omnibus Analysis. The final set of analyses were performed to resolve a conundrum that emerged consistently across four experiments. Specifically, we did not find that re-answer choices differed as a function of the number of models, even though single model problems were reliably more fluent than multiple model ones, and were also given higher FORs. This is in contrast to the item-based analysis, in which we showed that more fluent items were given higher FORs and were re-answered less often.

To solve the conundrum, we categorized all of the responses made in each experiment into one of four conditions: Single & Fluent, Single & Disfluent, Multiple & Fluent, and Multiple & Disfluent. That is, for each participant, we classified each of their fluent and disfluent responses according to whether it was given to a single-model or multiple-model. The number of responses in each condition are displayed in Table 7. As shown by the table, the majority of the responses to single-model syllogisms were fluent, and the multiple-model problems were less fluent, consistent with the main effect of model on Composite RT. However, is it also clear that within each model type, there was a lot of variability, in that a large plurality of responses to the single-model problems were disfluent, and many responses to the multiple-model problems were fluent. This variability may have been sufficient to mask the effect of model on re-answer choices. That is, although people chose to re-answer more often when they had responded disfluently, the fact that there were so many disfluent responses to single-model problems and fluent responses to multiple-model problems meant that the overall effect of number of models on re-answer choices was not observed.

Table 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
Total	2177	1722	1747	2142

General Discussion

The goal of these studies was to answer two questions: 1) is manipulating FOR ratings sufficient to affect re-answer choices? And 2) is the relationship between FOR and re-answer choices mediated by fluency? To answer these questions, we examined two variables: The anchoring variable, which did not affect answer fluency, and the number of models, which did. We also looked directly at the relationship between fluent and disfluent responses, defined for each participant, and their re-answer choices.

With respect to the first question, we have a clear answer: In four experiments, anchor values affected FOR ratings, but had no effect on re-answer choices. They also did not affect fluency. These data are inconsistent with Hypothesis A (Figure 1), and suggest that variables that raise or lower FOR judgments without also affecting the experience of fluency do not affect re-answer choices.

Instead, our data support Hypothesis B, and suggest that the relationship between FOR and re-answer choices is mediated by fluency. In four experiments, we observed that answers that are given fluently are less likely to be re-answered than those given less fluently. An important caveat to this conclusion is that a variable that reliably affected fluency, namely the number of models, did not reliably translate into an effect on re-answer choices. Our explanation for this disconnect is that the relationship between number of models and fluency was weak, as described in Table 7. Because the relationship between fluency and re-answer choices is reliable but not perfect, it is possible we failed to detect the potential relationship between models and re-answer choices.

Our data are consistent with the cue-utlization framework (Koriat, 2007), which suggests that monitoring judgements are based on experiences associated with solving the task or problem at hand. These would include, but are not limited to, feelings of fluency and disfluency. It is important to note that we do not claim that fluency is the only cue that people may rely on to inform re-answer choices, but it does appear that artificially inflating or lowering judgments via a manipulation like anchoring will not have that effect.

In this respect, it appears that metareasoning judgments are different than metamemory judgments. Recall that Yang et al. (2018) found that the anchoring manipulation affected both JOLs and re-answer choice. This is not the first occasion in which a discrepancy has emerged between metamemory and metareasoning. Thompson, Prowse Turner et al. (2013) observed that a cue that reliably affects JOL's, namely whether the problems were presented in a fluent or disfluent font, did not impact FOR judgments. That is, when asked to memorize words, participants reliably give higher JOLs to words printed in large than small font (e.g., Rhodes & Castell, 2008). Disfluent fonts, however, do not appear to have the same impact on FOR judgments. Thus, although both JOLs and FORs share some common cues, such as fluency, it is clear that they also draw on unique sources.

Another explanation for the discrepancy between our findings and Yang et al.'s (2018), is that the re-answer question may mean different things in the context of syllogistic reasoning and learning. Syllogistic reasoning is a difficult task, and participants may not have wanted to prolong exposure to the task by choosing to re-answer a lot of questions. Thus, they may have needed a higher FOR threshold to make a re-study choice than the comparable situation in learning words. Indeed, we note that participants chose to re-answer only between 20 and 30% of the trials, even though accuracy in some conditions was at chance levels.

For this reason, it would be desirable to replicate our findings using Thompson et al.'s (2011) two-response paradigm. In this paradigm, reasoners make a quick, intuitive response, and are then given a second opportunity to reflect on their answer. Strong FORs are associated with shorter rethinking times and fewer answer changes. It thus measures rethinking behaviour, rather than rethinking intentions. Given the concerns above, it may be a better measure of analytic engagement than re-answer choices on this task.

Finally, the data strongly suggest that asking people to make FORs does not change the way they approach the task. These data jibe with Thompson et al.'s 2011 and 2013 findings. In those papers, they compared responses generated using a two-response paradigm to those using a single-response paradigm, and found performance to be comparable in the two conditions. In the current paper, we found that the relationship between fluency and re-answer choices was similar, regardless of whether we asked participants for FOR judgments. Thus, we can be confident that this relationship is not caused by asking participants to reflect and make FOR judgments.

Conclusions

At the level of individual participants, there is a robust relationship between FORs, fluency, and re-answer choices. Responses that participants generate fluently engender higher FORs and lower re-answer choices. However, this relationship does not translate neatly to variables that affect fluency and FORs. Participants were more

fluent and gave higher FORs to single-than multiple-model problems, but this did not translate into re-answer choices. We speculate that this was due to the fact that a large plurality of single-model problems were answered disfluently, and that a large plurality of multiple-model problems were answered fluently, which might have masked the relationship between fluency and re-answer choices. We also observed that it is possible to affect FOR judgments without affecting fluency by means of an anchoring manipulation. This manipulation also did not translate into re-answer choices. Thus, we conclude that fluency experienced at the time of solving a problem produces higher FOR and re-study choices, but in order for variables that manipulate FOR to translate to re-study choices, they have to have a very strong effect on fluency.

Acknowledgments

We would like to thank Jamie Campbell and Carla Krachun for their helpful comments on earlier drafts of this paper.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*(8), 608-617. 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, in press). 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., & 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., & Narens, L. (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., Prowse 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., Prowse 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

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

Received: March 16, 2019

Appendix A

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.

Appendix B: Analysis of Accuracy for Experiment 1

The overall mean accuracy was 0.62. Data are plotted in Figure B1. Mean accuracy for valid syllogisms was higher (M = 0.71, SD = 0.02) than for their invalid counterparts (M = 0.53, SD = 0.02), $F(1,63) = 36.887$, $p < .001$, $\eta_p^2 = .369$. There was a significant interaction between model and validity, $F(1,63) = 10.195$, $p = .002$, $\eta_p^2 = .139$. These values are presented in Table B1. When the syllogisms were valid, participants were more accurate for single-model than for multiple-model problems (+0.09; $t(63) = 2.846$, $p = .006$), but this difference was absent for invalid syllogisms (-0.04; $t(63) = -1.309$, $p = .195$). The anchoring effect on accuracy was non-significant, $F(1,63) = 0.610$, $p = .438$, $\eta_p^2 = .010$.

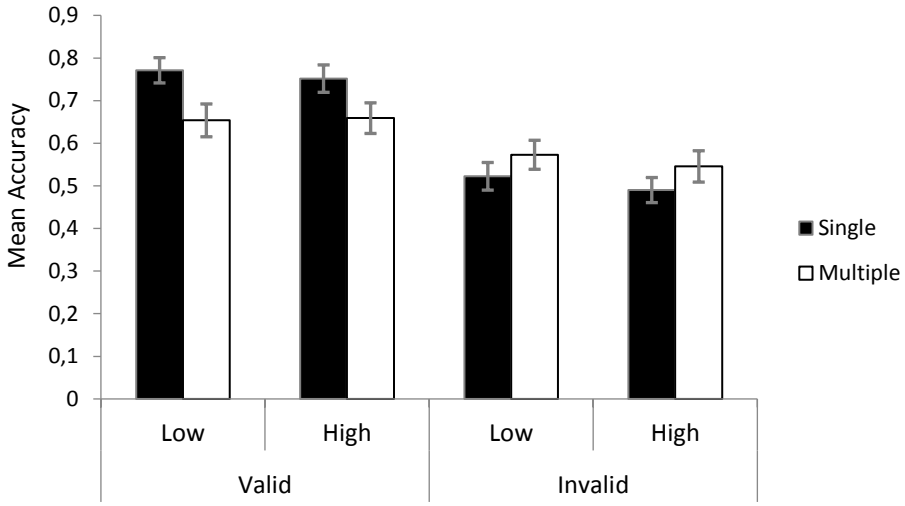


Figure B1. Mean accuracy for Experiment 1 as a function of anchor, model and validity. Error bars represent standard errors.

Table B1

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

Appendix C: Analysis of Accuracy for Experiment 2

Accuracy The overall mean accuracy was 0.63. The data for accuracy are plotted in Figure C1. 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), $F(1,63) = 13.805$, $p < .001$, $\eta_p^2 = .180$. There was also a main effect of validity, $F(1,63) = 12.716$, $p = .001$, $\eta_p^2 = .168$. 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, $F(1,63) = 34.986$, $p < .001$, $\eta_p^2 = .357$. These values are presented in Table C1.When the syllogisms were valid, single-model problems were answered more accurately than multiple-model problems (+0.19, $t(63) = 7.994$, $p < .001$), but the difference was not present when the syllogisms were invalid (-0.04, $t(63) = -1.447$, $p = .153$). 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, $t(63) = 2.533$, $p = .014$), but the difference was smaller for problems paired with low anchors (+0.06, $t(63) = 2.272$, $p = .026$).

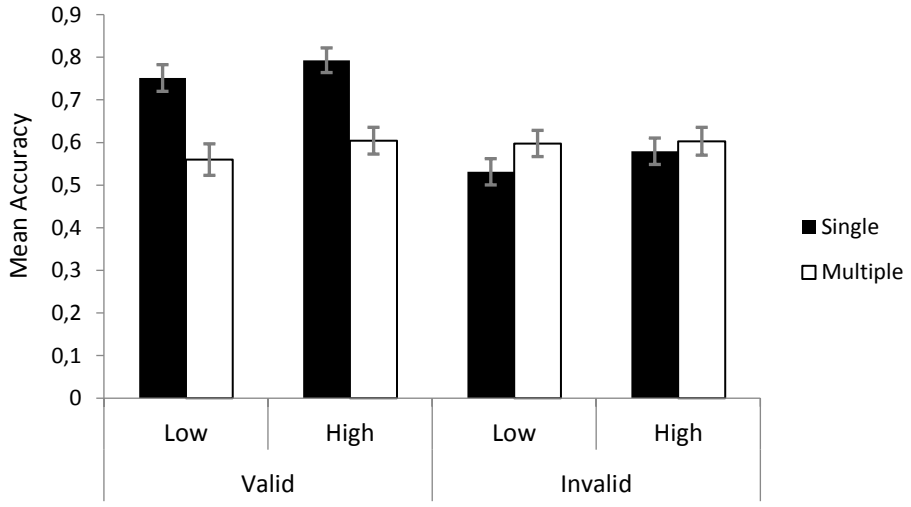


Figure C1. Mean accuracy in Experiment 2 as a function of model, anchor, and validity. Error bars represent standard errors.

[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 C1

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

Appendix D: Analysis of Accuracy Data for Experiment 3

The overall mean accuracy was 0.61. Data are plotted in Figure D1. 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), $F(1,63) = 21.702$, $p < .001$, $\eta_p^2 = .256$. Similar to the previous experiments, the interaction between model and validity was also significant, $F(1,63) = 37.487$, $p < .001$, $\eta_p^2 = .373$. The values of the interaction are displayed in Table D1. When the problems were valid, participants were more accurate on the single-model syllogisms (+0.14, $t(63) = 5.391$, $p < .001$), whereas their accuracy was higher on the multiple-model syllogisms when the problems were invalid (-0.09, $t(63) = -2.871$, $p = .006$). Consistent with Experiment 1, the main effect of model on accuracy was non-significant, $F(1,63) = 1.969$, $p = .165$, $\eta_p^2 = .03$.

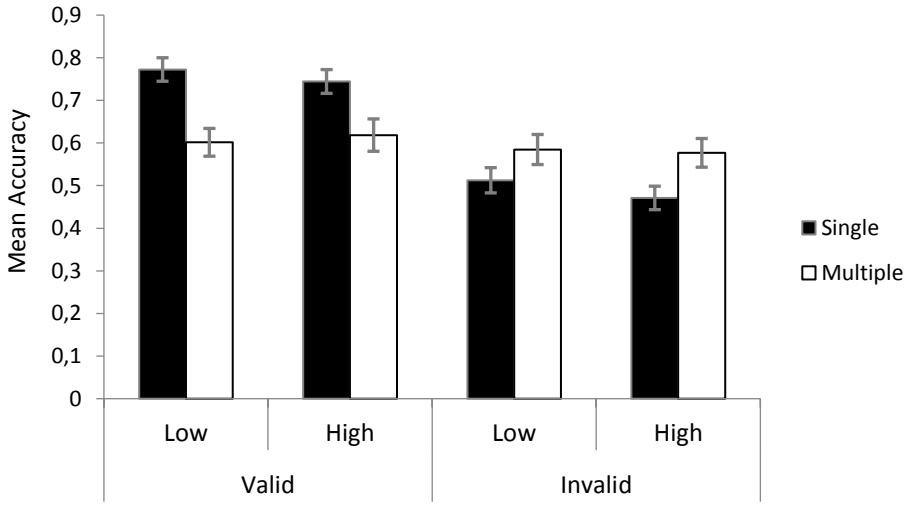


Figure D1. Mean accuracy in Experiment 3 as a function of model, anchor, and validity. Error bars represent standard errors.

Table D1

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

Appendix E: Analysis of Accuracy Data for Experiment 4

The overall mean accuracy was 0.66. Data are plotted in Figure E1. Consistent with previous experiments, there was a main effect of validity on accuracy, $F(1,63) = 29.807$, $p < .001$, $\eta_p^2 = .321$. 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, $F(1,63) = 37.103$, $p < .001$, $\eta_p^2 = .371$. These values are presented in Table E1. Participants correctly answered more single-model problems than multiple-model ones when the syllogisms were valid (+0.14, $t(63) = 5.325$, $p < .001$), but they were more accurate on multiple-model problems than their single-model counterparts for the invalid problems (-0.09, $t(63) = -3.541$, $p = .001$).

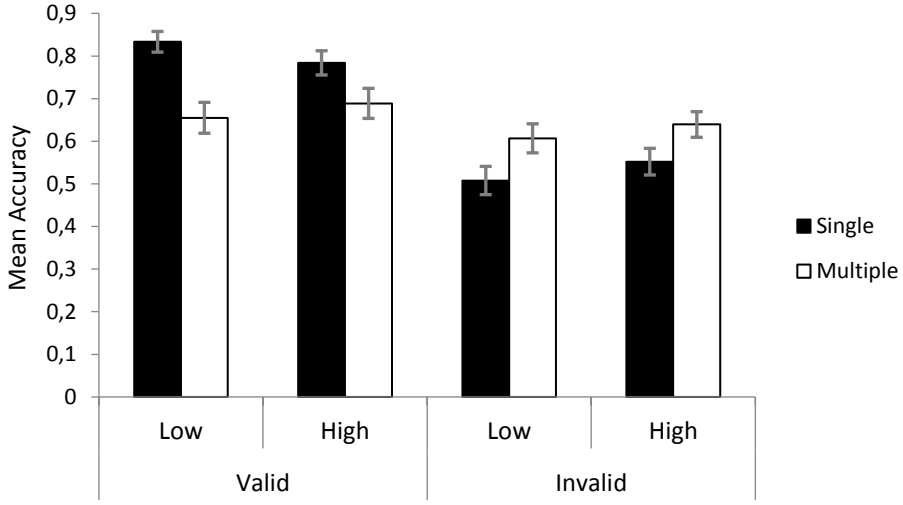


Figure E1. Mean accuracy as a function of model, anchor, and validity. Error bars represent standard errors.

Table E1

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