Intuition, reason, and metacognition

Valerie A. Thompson *, Jamie A. Prowse Turner, Gordon Pennycook

Department of Psychology, University of Saskatchewan, SK, Canada S7N 5A5

A R T I C L E   I N F O

Article history:
Accepted 31 May 2011
Available online 27 July 2011

Keywords:
Metacognition
Reasoning
Dual Process Theories
Intuition
Analytic thinking
Retrospective confidence

A B S T R A C T

Dual Process Theories (DPT) of reasoning posit that judgments are mediated by both fast, automatic processes and more deliberate, analytic ones. A critical, but unanswered question concerns the issue of monitoring and control: When do reasoners rely on the first, intuitive output and when do they engage more effortful thinking? We hypothesised that initial, intuitive answers are accompanied by a metacognitive experience, called the Feeling of Rightness (FOR), which can signal when additional analysis is needed. In separate experiments, reasoners completed one of four tasks: conditional reasoning ($N = 60$), a three-term variant of conditional reasoning ($N = 48$), problems used to measure base rate neglect ($N = 128$), or a syllogistic reasoning task ($N = 64$). For each task, participants were instructed to provide an initial, intuitive response to the problem along with an assessment of the rightness of that answer (FOR). They were then allowed as much time as needed to reconsider their initial answer and provide a final answer. In each experiment, we observed a robust relationship between the FOR and two measures of analytic thinking: low FOR was associated with longer rethinking times and an increased probability of answer change. In turn, FOR judgments were consistently predicted by the fluency with which the initial answer was produced, providing a link to the wider literature on metamemory. These data support a model in which a metacognitive judgment about a first, initial model determines the extent of analytic engagement.

© 2011 Elsevier Inc. All rights reserved.
1. Introduction

There is much evidence to support the thesis that reasoning and decision-making are accomplished by recourse to two qualitatively different types of processes (see Evans and Frankish (2009) for a recent review), differing in terms of the degree to which they are characterized as fast and automatic (Type 1) or slow and deliberate (Type 2). A variety of Dual Process Theories (DPT) have been proposed to explain the interaction of these two processing systems (e.g., Evans, 2006; Kahneman, 2003; Sloman, 2002; Stanovich, 2004). Although they make somewhat different claims about the extent, degree, and timing of Type 2 processes, they share the basic assumption that automatic Type 1 processes give rise to a highly contextualised representation of the problem and attendant judgments that may or may not be analysed extensively by more deliberate, decontextualised Type 2 processes.

According to DPTs, the outcome of a given reasoning attempt is determined jointly by the content of the information that is retrieved by Type 1 processes (see Kahneman (2003) and Stanovich (2004) for extensive analyses) and by the quality and depth of Type 2 processing. As such, the explanatory value of DPTs depend critically on their ability to predict the circumstances under which Type 2 processes are more or less engaged (Evans, 2009; Stanovich, 2009; Thompson, 2009, 2010). To date, explanations have focussed on global characteristics of the reasoner, such as cognitive capacity (De Neys, 2006a; Stanovich, 1999) or aspects of the environment, such as the amount of time allotted to complete the task (Evans & Curtis-Holmes, 2005; Finucane, Alhakami, Slovic, & Johnson, 2000), the instructions provided to the reasoner (Daniel & Klaczynski, 2006; Evans, Newstead, Allen, & Pollard, 1994; Newstead, Pollard, & Evans, 1992; Vadeboncoeur & Markovits, 1999), or variables that create a global perception of difficulty, such as presenting problems in a difficult-to-read font (Alter, Oppenheimer, Epley, & Eyre, 2007).

Missing from this analysis is an account of item-specific cues that trigger Type 2 thinking. To illustrate, consider the following two items. One is taken from Frederick’s (2005) Cognitive Reflection Test (CRT) and the second is an isomorphic version of it (Thompson, 2009):

If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets?
___ minutes

If it takes 5 machines 2 min to make 10 widgets, how long would it take 100 machines to make 100 widgets?
___ minutes

The first problem strongly cues the response “100”, which is, in fact, erroneous but often given as an answer (Frederick, 2005). From a DPT view, Type 1 processes produce an initial response to the first version of the problem (i.e., 100). This answer is then examined by Type 2 processes, determined to be satisfactory, and (incorrectly) given as the answer by a large majority of participants (Evans, 2009; Kahneman, 2003). Less clear is the explanation for why this answer is so readily deemed to be satisfactory and the subsequent Type 2 analysis is so cursory; also missing is the explanation for why the second version of the problem is more likely to suggest that mental effort (Type 2 processing) will be needed to achieve the solution.

Such variability in performance across nominally equivalent problems is common (e.g. Buccarelli & Johnson-Laird, 1999; Johnson-Laird, 1983; Marcus & Rips, 1979). The question, therefore, becomes this: For a given participant of a given cognitive capacity, operating under a given set of task instructions, in a given environment, what predicts the degree of Type 2 engagement? In the current paper, we propose an answer to this question that is grounded in basic metacognitive processes. Specifically, we posit that a third category of process monitors Type 1 outputs (Simmons & Nelson, 2006; Thompson, 2009, 2010) and determines the extent of Type 2 engagement (see Evans (2009) and Stanovich (2009) for related discussions).

1.1. The Metacognitive Reasoning Theory: the Feeling of Rightness and Type 2 thinking

This proposal draws heavily on the metamemory literature, which has long acknowledged the distinction between the processes responsible for retrieving information from memory and the processes
responsible for monitoring that information (see Dunlosky and Bjork (2008) for an overview). Monitoring refers to the “subjective assessment of one’s own cognitive processes and knowledge” (Koriat, Ma’ayan, & Nussinson, 2006, p. 38). This assessment can be derived inferentially from implicit cues, such as the ease with which a memory comes to mind (Benjamin, Bjork, & Schwartz, 1998; Koriat & Ma’ayan, 2005), or based on explicit cues, such as beliefs about one’s skill at a task (e.g., Dunning, Johnson, Ehrlinger, & Kruger, 2003; Prowse Turner & Thompson, 2009; see Koriat (2007), for a review).

By extension, we posit a set of processes that monitor Type 1 outputs, which in turn, determine the depth of Type 2 thinking. Under this proposal, Type 1 processes generate two distinct outputs: The first is the content of the initial answer and the second is an accompanying sense of the correctness of that answer (Simmons & Nelson, 2006; Thompson, 2009). This Feelings of Rightness (FOR) is predicted to vary in strength across a set of problems. On problems such as the first widget problem above, the FOR should provide a compelling cue that the initial inference is the correct one. In a variety of other situations, the FOR will be less compelling. In this way, the generation of Type 1 outputs is proposed to be analogous to memory retrievals (Thompson, 2009, 2010), which, in addition to a specific memory content, also carry an affective component that acts as a cue to the correctness of the retrieval (e.g., Fazendeiro, Winkielman, Luo, & Lorah, 2005; Gruppuso, Lindsay, & Masson, 2007; Johnston, Dark, & Jacoby, 1985; Whittlesea, 1993; see Koriat (2007) for an extensive review).

Crucially, from the point of view of the current argument, the metacognitive experiences associated with memory retrievals determine both the allocation of resources as well as study or problem solving strategies. For example, metacognitive experiences predict the length of time spent studying an item for a subsequent recall test (e.g., Mazzoni & Cornoldi, 1993; Nelson, 1993; Son, 2004; Son & Metcalfe, 2000), the amount of time spent searching for an item in memory (Singer & Tiede, 2008), the decision to mass or distribute practice (Benjamin & Bird, 2006; Son, 2004), as well as the decision to derive a solution by computation vs. retrieve the answer from memory (Reder & Ritter, 1992). By analogy, therefore, the Feeling of Rightness (FOR) that accompanies Type 1 processing should signal whether the current output suffices or whether additional Type 2 processes are needed (Thompson, 2009, 2010).

Surprisingly, there is little research on the role of these types of metacognitive processes in reasoning and decision-making. That is, the monitoring and control aspects of metacognitive function have been relatively neglected in this field, despite the large literature on the factors that produce overconfidence in judgments (e.g., Caputo & Dunning, 2005; Dunning et al., 2003; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Hansson, Juslin, & Winman, 2008).

One exception is a study by Simmons and Nelson (2006), who provided preliminary support for a monitoring mechanism that constrains Type 2 analysis. Their task required participants to place a wager “against the spread”. This required them to decide whether the favoured team would win by a given margin, or whether the underdog would lose by less than that margin. Although the spread is constructed to equalise the probability of the two outcomes, people overwhelmingly bet on the favourite. Thus, betting on the favourite was categorized as the intuitive response. Confidence in the intuitive response was operationalised as the degree of consensus that the favourite would win, regardless of the spread. They found that the higher the degree of consensus, the more likely participants were to wager on the favourite, and the more confident they were in their bet. On this basis, they concluded that intuitive answers are often accompanied by a sense of correctness, which then determines the probability of subsequent processing.

The experiments reported in the current paper advanced on this initial work in a number of ways: First, we sought to measure the FOR explicitly, rather than by using a proxy such as consensus. This allowed us to make a direct link to the metamemory literature, where metacognitive constructs such as the Feeling of Knowing or Judgments of Learning are solicited directly from participants (see Van Overschelde (2008), for a recent overview). The direct measurement also allowed us to rule out an alternative explanation for the relationship between the consensus variable and the measure of intuitive responding, namely that both reflected a preference to choose the favourite. Importantly, we also derived measures of Type 2 engagement apart from participants’ tendency to give the intuitive response. This is because giving the intuitive answer in no way implies the absence of Type 2 thinking, given that participants may decide, after a lengthy period of Type 2 thinking, to go with the intuitive...
answer after all. Finally, we tested a number of hypotheses about the underlying determinants of the FOR.

1.2. Experimental design and predictions

1.2.1. Two response paradigm

Our first goal was to examine how participants monitor and regulate their performance over a series of problems. To do so, we modified Koriat and Goldsmith’s (1996) quantity-accuracy profile (QAP; see Goldsmith and Koriat (2008) for a recent review). Their procedure was designed to test, among other things, the efficacy of monitoring initial retrievals to determine whether they should be subsequently given as a response. For the QAP, participants are given two memory tests. On the first, they are required to respond to every item; they then give a confidence judgment, which is used to predict their performance under free-report conditions.

We adapted this procedure to fit the current requirements as follows: Participants gave two responses to each of a series of reasoning problems. For the first, participants were told that we were interested in studying reasoners’ intuitions and they were instructed to give the first answer that came to mind. Following this initial response, a subjective measure of the FOR was taken using a likert scale. The format of the scale varied somewhat; in some cases they were asked to evaluate certainty and in others, rightness. They were told to give the answer that was their first instinct or gut feeling. As a manipulation check, we asked them to indicate whether or not they had, indeed, done so for each trial. This initial response presumably reflects the outcome of Type 1 processing with minimal Type 2 analysis.

This assumption is based on the findings of several studies indicating that fast responses are more likely than slow responses to reflect the output of heuristic, Type 1 processes (De Neys, 2006b; Evans & Curtis-Holmes, 2005; Finucane et al., 2000; Roberts & Newton, 2001; Tsuji & Watanabe, 2010). For example, when forced to respond quickly, reasoners are more likely to respond on the basis of conclusion believability than when allowed additional time to consider their responses (Evans & Curtis-Holmes, 2005; Tsuji & Watanabe, 2010); they are also more likely to show matching bias on Wason’s selection task (Roberts & Newton, 2001) and to make choices guided by affect (Finucane et al., 2000). Similarly, heuristic, Type 1 responses require less time to produce than their analytic counterparts (De Neys, 2006b). Finally, near-infrared spectroscopy shows that the interior frontal cortex, which is involved in inhibiting belief-based responses, is less engaged when participants are forced to respond quickly than when they are not (Tsuji & Watanabe, 2010). Thus, requiring a fast response from participants should produce responses that are based largely on the output of Type 1 processes.

1.2.2. Measures of Type 2 processes

To measure Type 2 engagement, participants were allowed as much time as needed to produce a final answer to the problems. Although the instructions were tailored to the specific tasks that participants completed, they all indicated that participants should be sure at this point that they had taken their time and thought about the problem carefully.

From this, we derived three measures of Type 2 engagement. The first measure was the degree or probability of change from the first answer to the second answer. A change of answer would indicate that some additional analysis had taken place and should therefore be a reliable index of Type 2 thinking. Note that while a change of answer should reliably indicate Type 2 engagement, failure to do so is not evidence for the absence of Type 2 engagement. That is, there is reason to believe that at least some Type 2 thinking is directed at rationalising the initial response (e.g., Evans, 1996; Shynkaruk & Thompson, 2006; Stanovich, 2004, 2009; see also Wilson and Dunn (2004) for a related discussion). Thus, we also measured the amount of time spent re-thinking each problem. Given that Type 2 processes are assumed to be deliberate, time consuming processes, the amount of time spent engaging in a problem should be a reliable index of the extent of Type 2 processing (De Neys, 2006b).

Our third measure was a traditional measure of analytic engagement, namely whether or not the final answer was correct by a relevant normative standard. Such measures are presumed to reflect successful application of the rules of probability or logic; success by these standards is typically more
likely among those of high cognitive or working memory capacity (De Neys, Schaeken, & d’Ydewalle, 2005a; De Neys & Verschueren, 2006; Stanovich, 1999) and thus thought to be a signature of capacity-demanding, deliberate Type 2 processes. However, given that Type 2 processes may also be engaged to produce normatively incorrect responses (Evans, 2007b; Stanovich, 2009) and that normatively correct responses may be produced by non-analytic processes (Gigerenzer, Todd, & the ABC Reasoning Group, 1999; Oaksford & Chater, 2007), this latter measure should be considered the least reliable indicator of Type 2 engagement.

1.2.3. Design and predictions

The hypothesis tested was that initial responses to reasoning problems are accompanied by a FOR that determines the quality and extent of Type 2 processes. On this basis, it was predicted that low FOR’s should lead to longer rethinking times, greater probability of answer change, and increased probability of normatively correct responding than high FOR’s.

There were two strategies for testing the hypothesis. In all four experiments, we measured FORs across a series of trials and used this to predict rethinking times, answer changes, and the probability of giving a normatively correct answer. In Experiments 3 and 4, we manipulated factors that were predicted to affect FOR judgments, and compared outcomes in conditions with relatively lower and higher FOR judgments.

We used four different reasoning tasks: Experiment 1 tested conditional reasoning, a form of deductive reasoning in which participants are asked to make inferences about conditional relationships (e.g., if the car runs out of gas, then it stalls). Experiment 2 tested a more complex version of this task in which participant were asked to determine what follows from two conditional statements (e.g., if something is a rose, then it has a gebber; if something has a gebber, then it is a flower). Experiment 3 used a probability judgment task, in which participants were asked to estimate the probability than an individual belonged to a category given two pieces of information: the base rate of the category and a personality description. Finally, Experiment 4 used categorical syllogisms, in which participants were asked to evaluate the validity of conclusions drawn from quantified premises.

1.3. Determinants of the Feeling of Rightness

In addition to testing the relationship between FORs and Type 2 thinking, a second goal was to investigate potential determinants of the FOR. In the metamemory literature, several variables have been demonstrated to mediate metamemory judgments; many of these variables are tied to retrieval processes, rather than to the contents of memory per se (e.g., Benjamin et al., 1998; Busey, Tunnicliff, Loftus, & Loftus, 2000; Jacoby, Kelley, & Dywan, 1989; Koriat, 1995, 1997; Koriat & Levy-Sadot, 1999; Schwartz, Benjamin, & Bjork, 1997). For example, familiarity of the retrieval cues, as opposed to familiarity of the answer determines Feelings of Knowing (Reder & Ritter, 1992; Schunn, Reder, Nhouyvanisvong, Richards, & Stroffolino, 1997; Vernon & Usher, 2003), as does the amount of ancillary information that is brought to mind during the retrieval attempt (Koriat, 1993, 1995; Koriat, Levy-Sadot, Edry, & de Marcas, 2003).

We identified three variables that might play a similar role in reasoning judgements; one was a task-independent variable called answer fluency which was tested in all four experiments. In addition, we examined two task-specific variables, namely the probability that a conclusion is accepted as valid (Experiments 1 and 2) and the presence or absence of competing responses (Experiment 3).

1.3.1. Fluency and the FOR

Answer fluency refers to the ease with which the initial conclusion comes to mind. In the metamemory literature, there is much evidence to suggest that the fluency with which items can be retrieved from memory is a powerful determinant of the sense that they have been or will be accurately remembered (e.g., Benjamin et al., 1998; Jacoby et al., 1989; Kelley & Jacoby, 1993, 1996; Matvey, Dunlosky, & Guttentag, 2001; Whittlesea & Leboe, 2003). Indeed, fluent processing can produce an illusion that an item has been previously experienced, regardless of whether it has or not (e.g., Jacoby et al., 1989; Whittlesea, Jacoby, & Girard, 1990). Similarly, and critical for the current work, subjective confidence in the correctness of a memory retrieval varies as a function of the speed with which the answer
comes to mind (Costermans, Lories, & Ansay, 1992; Kelley & Lindsay, 1993; Robinson, Johnson, & Herndon, 1997). Consequently, the most straightforward prediction about the origins of the FOR accompanying an initial answer is that it is determined by the speed or fluency with which that answer is produced.

This concept of answer fluency is distinct from that of processing fluency (e.g., Koriat, 2007; Van Overschelde, 2008) that has been shown to affect other types of complex judgments (see Schwarz (2004) for a summary). Making a stimulus difficult to perceive can affect judgments of aesthetic pleasure (Reber, Schwarz, & Winkielman, 2004), truth (Reber & Schwarz, 1999), and reasoning (Alter et al., 2007). For example, Alter et al. (2007) demonstrated that reasoners were more likely to give normatively correct answers to problems like the first widget problem above when the problems were presented in a difficult to read font (disfluent) than an easy to read one (fluent). In other words, when the experience of processing a task was fluent, there was less Type 2 engagement relative to conditions where the experience was less fluent.

Although these data offer support for the hypothesis that fluency, broadly construed, may act as a cue to Type 2 processing, they do not speak to the processes that monitor outputs on a trial by trial basis. That is, although answer fluency and processing fluency are sometimes treated as interchangeable constructs, (Alter & Oppenheimer, 2009; Briñol, Petty, & Tormala, 2006), they are, in fact, distinct. Specifically, for a given level of processing fluency, answer fluency and the accompanying FOR might vary substantially across items for the simple reason that some answers will be produced faster than others. We propose these sources of variance are linked, such that answer fluency and the accompanying FOR can explain why some items elicit more or less Type 2 thinking than others. In all four Experiments, we measured variation in both fluency and FOR judgments for each trial, using the latter measure to predict the probability and extent of Type 2 engagement. In Experiments 3 and 4 we manipulated variables that were predicted to affect the fluency of responses, and then compared FOR judgments across conditions.

1.3.2. Other determinants of the FOR

The current paradigm also afforded us the opportunity to explore two other potential determinants of the FOR. In previous studies of deductive reasoning (used as paradigms in Experiments 1 and 2), we have noticed that reasoners accept a provided conclusion as valid more often than warranted by chance (or by other task-relevant features such as the validity or believability of the conclusion; Shynkaruk & Thompson, 2006). A possible explanation for this is that the decision to accept a conclusion is accompanied by a higher FOR than the decision to reject it; we will test this hypothesis in Experiments 1 and 2.

In addition, many reasoning theories give a special status to problems for which two or more answers come into conflict (e.g., Amsel et al., 2008; Ball, Phillips, Wade, & Quayle, 2006; Evans, 2006; Sloman, 1996), as for example, when a conclusion is valid but not believable. The need to resolve the conflict is posited to cue additional Type 2 processing that would not be engaged in its absence (De Neys & Glumicic, 2008; Evans, 2007a). In Experiment 3, we will test the hypothesis that FOR’s are sensitive to conflicting inputs, such that the presence of conflict lowers the FOR.

2. Experiment 1

The goal of this experiment was twofold. The first was to test the hypotheses generated from the Metacognitive Reasoning Theory regarding the relationships between FOR and Type 2 thinking; the second was to test the relationship between FOR and two putative determinants, namely answer fluency and probability of conclusion acceptance. These hypotheses were tested using the conditional inference task described below, modified to our new two-response procedure. We were not interested in performance on the task per se, which is well-documented (e.g., Cummins, Lubart, Alksnis, & Rist, 1991; De Neys, Schaeken, & d’Ydewalle, 2005b; Markovits & Nantel, 1989; Thompson, 1994, 2000), but rather, used the task as a vehicle to study metacognitive monitoring. Two sets of predictions were tested: (1) Answers accompanied by a low FOR should promote more Type 2 thinking relative to answers accompanied by a strong FOR and (2) the FOR, in turn, should be higher for fluently generated
answers than their less fluent counterparts and for conclusions that were accepted as opposed to rejected.

Conditional reasoning requires participants to make inferences about statements of the form “if p, then q” (see Evans & Over, 2004; Oaksford, 2010). The minor premise asserts the occurrence or non-occurrence of one of the propositions and then participants are asked what can be inferred about the other. This produces four inferences for evaluation, Modus Ponens (MP), Modus Tollens (MT), Affirming the Consequent (AC), and Denying the Antecedent (DA), of which the first two are logically valid, and the second two are not, e.g.:

If a car runs out of gas, then it will stall.

The car has run out of gas. Therefore it will stall. (MP: valid)
The car has not run out of gas. Therefore it will not stall. (DA: invalid)
The car has run out of gas. Therefore it will stall. (MT: valid)
The car has run out of gas. Therefore it will not stall. (DA: invalid)

The only difference between the task used in the current study and the usual version of it was that participants were asked to generate two responses to each problem using the two-response paradigm outlined previously. That is, they generated an initial, intuitive response, provided an FOR, and then were allowed as much time as needed to generate a final conclusion.

The second goal of this experiment was to establish that this modified procedure did not alter the outcome of participants’ reasoning processes. To this end, we included two control groups who gave only a single response to each problem. The first did so under free-time conditions; responses from this group were compared to the final responses made by the two-response group in order to ensure that the two-response procedure did not interfere with the answers participants gave. A second control group generated only an intuitive answer and FOR to each problem. These FOR’s were used to predict answer changes in the two-response group and thinking times in the two-response and free-response conditions. This allowed us to rule out the possibility that differences in thinking times and answer changes were produced by having participants make FOR judgments. Because the latter set of analyses required comparing responses across groups of participants using items as the units of analysis, we chose a set of materials for which many items were available (Thompson, 1994).

2.1. Method

2.1.1. Participants

Initially, 30 participants were recruited for each of the two-, free-, and intuitive-response groups. Before analysing the data, we did a manipulation check of the two-response and intuitive-response groups; these participants were asked to indicate on every trial whether or not they had complied with the instruction to give the first answer that came to mind. Three and two participants respectively were replaced because they indicated that they had complied on less than 90% of trials. The final sample of 90 participants had a mean age of 23.5 years ($sd = 6$ years); 60 (67%) were females. They were randomly assigned to the three groups. Participants were either paid CAN$5 for participation or received partial course credit; the former were members of the University community who responded to posted advertisements and the latter were University of Saskatchewan introductory psychology students.

2.1.2. Materials

The experiment was programmed and run using E-Prime (Schneider, Eschman, & Zuccolotto, 2002) and presented on a high-resolution computer monitor. On each trial, participants were shown a conditional statement and asked to evaluate the validity of one of four inferences: Modus Ponens (MP), Modus Tollens (MT), Affirming the Consequent (AC), and Denying the Antecedent (DA).

Sixteen conditional sentences of the form “If p then q” were taken from Thompson (1994). All of these described typical, everyday events. The inferences varied orthogonally with respect to their validity and believability. Thus, regardless of which criteria was used to evaluate the inferences, we would expect participants to accept and reject equal numbers of inferences.
The believability of the inferences was manipulated by changing the relationship of the $p$ and $q$ propositions. To the extent that occurrence of $p$ is sufficient to bring about the occurrence of $q$, as in the “car” example above, the valid inferences are believable. When $p$ is not sufficient for $q$, the valid inferences are less believable (e.g., if the TV is plugged in, then it works; the TV is plugged in, therefore, it works). Similarly, to the extent that $p$ is a necessary condition to bring about $q$, the invalid inferences are believable (e.g., if the TV is plugged in, then it works; the TV works, therefore, it is plugged in).

Eight sentences described causal relationships, like the car and TV examples above; eight described categorical relationships, e.g., If an animal is a robin, then it is a bird. Each conditional statement was presented four times, once with each inference type (MP, MT, AC, and DA), resulting in a total of 64 problems. Half of the inferences were believable and half were unbelievable.

In addition to these 64 problems, two additional problems were constructed to serve as practice problems. These practice problems were also taken from Thompson (1994), but differed from the 16 test sentences. Both were believable inferences: one was a causal MP problem and the other was a categorical DA problem.

2.2. Procedure

Participants were tested individually and each evaluated the same 66 conditional statements. Testing took approximately 30 min. The problems appeared on the monitor one at a time. The first two conditional statements were the practice problems, and were presented in random order. They were included to ensure that participants were familiar with these types of questions before data was recorded. The remaining 64 test problems were each presented once, in a different random order for each participant.

Before beginning the task, all participants were provided with a set of general instructions presented on the computer screen. These instructions explained the nature of the reasoning task, gave an example, explained how participants were to enter their responses, and warned that the same content would be associated with several different reasoning problems. Participants were instructed to choose “yes” if they judged that a conclusion followed logically from the information given and “no” if they did not. The response options, “yes” and “no”, appeared below each problem and participants selected their response by pressing either the 1 or the 3 on the keyboard. Assignment of “yes” and “no” to the 1 and 3 keys was counterbalanced so that for half of the participants 1 corresponded to “yes” and for the other half, 3 corresponded to “yes.”

Following each response, participants were provided with the following scale on which they were to express their Feeling of Rightness in their answer by typing in the appropriate number on the keyboard:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Feeling</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>Certain</td>
<td>Fairly</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>I'm right</td>
<td>Certain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Whereas participants in the free-response group gave a single response with no time constraints, participants in the two-response conditions were told that they would give two responses to each problem. They were told to provide their first answer as soon as they had read the problem and to give the first answer that came to mind. The importance of responding quickly was stressed, as was the fact that they would be timed. They were told that they would be allowed to give a final answer, and that they should be sure that this one was logically valid. After this initial set of instructions, they then completed the practice problems, and prior to the test trials, these instructions were reiterated. Specifically, the need for intuitive responding on the first response was re-emphasised and participants were instructed not to second guess their first answer; they were also told to ensure that their second answer was correct. Participants in the intuitive-response group also gave only one response under instructions identical to the first-response condition described above.
Timing began when the problem appeared and ended when the participant entered their first response. After making their first response and FOR and before giving their final answer, they were asked “did you give your first response to the last reasoning problem?” The problem then re-appeared and participants were allowed as much time as they wanted to give their second answer. This rethinking-period was also timed, starting when the problem re-appeared on the screen and ending when participants entered their answer. Finally, participants made a final judgment of confidence (FJC) using the scale presented above.

2.3. Results

2.3.1. Scoring

A conclusion was scored as “accepted” when participants responded “yes” (i.e., indicated it followed from the premises). A response was scored as normatively accurate when participants responded “yes” to valid and “no” to invalid conclusions. All RT measures were converted to log10 prior to analysis.

2.3.2. Analysis strategy

The data are reported below in several sections. The first data reported are those that test the hypotheses about the relationship between FOR, Type 2 engagement, answer fluency, and probability correct. Each of the subsequent analyses diffuses potential counter-arguments to the conclusions reached on the basis of these initial findings, including potential arguments about our success in eliciting participants’ intuitive answers and the potential for the first judgment to affect the outcome of the second. These item-based analyses compared the two-response group to the intuitive- and free-response groups. In the final section, we present a quick summary of the FJC data.

2.3.3. Relationship between FOR and three measures of Type 2 engagement

In this section, we tested the hypothesis that initial responses that were accompanied by a low FOR would be associated with more Type 2 thinking than responses associated with a high FOR. To do this, we examined the relationship between the FOR at Time 1 and three measures of Type 2 engagement: The amount of time spent re-thinking a problem, the probability that the answer changed from Time 1 to Time 2, and the probability of providing a normatively correct answer at Time 2. The analyses reported here pertain only to the two-response group.

Three predictions were tested: (1) for a given participant, items that received more Type 2 engagement should have been given lower FORs than items that received high FORs; (2) for a given item, participants who gave low FORs should have engaged more Type 2 thinking than participants who gave high FORs (this also served as a control for item difficulty effects as each item was analysed separately); (3) across participants, items that were given low FORs should be characterised by more Type 2 thinking than those given high FORs.

2.3.4. FOR and Type 2 engagement: participant analysis

For each participant, we identified items on which they changed their answers and compared the mean FOR to items on which their answers did not change (participants changed their answers 13% of the time). FOR was lower when participants changed answers than when they did not, t(26) = 4.64, p < .001.1 To determine the relationship between rethinking time and FOR, we computed each participant’s correlation between FOR and rethinking time, then computed the average correlation across participants. As expected, the correlation was negative and significantly different from zero (Mr = −.31, t(26) = 8.51, p < .001), indicating that stronger FOR judgments were associated with shorter periods of rethinking. These data support the predictions of the Metacognitive Reasoning Theory regarding the relationship between FOR and analytic engagement: Low FOR judgments were associated with longer rethinking times and higher probability of changing answers, whereas high FOR judgments were associated with shorter rethinking times and a lower probability of changing answers.

1 Note that three participants did not change any responses and so were excluded from the analysis.
The relationship between FOR and normative accuracy was less straightforward. On the assumption that normative accuracy is a by-product of analytic engagement, we predicted that the FOR should be negatively correlated with normative accuracy. Instead, the FOR for answers that were correct at Time 2 (\(M = 5.37, sd = .90\)) was higher than the FOR for incorrect answers (\(M = 5.22, sd = .93\)), \(t(29) = 2.06, p = .049\). As argued earlier, normative accuracy should be considered the least reliable indicator of analytic engagement; indeed, in the following section, we observe that longer rethinking times were associated with less accurate responses. Instead, the positive relationship suggests that the FOR judgments are calibrated with item difficulty, such that more difficult items are given lower ratings than easier ones (Koriat, 2008).

2.3.5. FOR and Type 2 engagement: per item analyses

This raises an alternative interpretation of the relationship between FOR, rethinking times, and answer changes, namely that difficult items took longer to complete, were more likely to be changed, and given lower FORs than easier ones. To rule out an explanation based on item difficulty, for each item, we computed a correlation between the 30 participants’ FOR ratings, a binary code indicating whether their answer had changed, and the amount of time spent rethinking. When averaged across items, both the mean correlation between FOR and RT (\(M = -.34, sd = .16\)) and the mean correlation between FOR and probability of change (\(M = -.21, sd = .18\)) were less than zero on a one-sample \(t\)-test, \(t(63) = 16.8\) and \(t(61) = 9.14, p < .001\). In other words, the correlation between FOR and the measures of Type 2 engagement remained robust, even after controlling for item difficulty.

2.3.6. FOR and Type 2 engagement: item-based correlations

In this section, we replicated the relationship between FOR and Type 2 engagement using items as the units of analysis. For this, we computed, for each of the 64 items, the mean FOR (\(M = 5.30, sd = .34\)), the mean time spent rethinking (\(M = 6.21\) s, \(sd = 1.64\)), the mean number of participants who changed their answers for that item (\(M = 4.03, sd = 2.21\)) and the accuracy of responses at Time 2 (\(M = .59, sd = .26\)). The correlation matrix is presented in Table 1.

As predicted, there were large, negative correlations between FOR, the amount of time spent rethinking, and the probability that an answer changed during the rethinking period. That is, participants spent longer thinking about and were more likely to change their minds about items that engendered a low FOR than those for which the FOR was high. These data are consistent with the Metacognitive Reasoning Theory, in which an initial answer to a problem is accompanied by a metacognitive experience, which, in turn, predicts the extent of Type 2 processing.

Again, the relationship to normative accuracy is less straightforward. Consistent with the participant-based analyses, the correlation between FOR and normative accuracy was positive, rather than

<table>
<thead>
<tr>
<th></th>
<th>FOR</th>
<th>RT-1st</th>
<th>Accept-1st</th>
<th>Answer change</th>
<th>Rethink time</th>
<th>Accept-2nd</th>
<th>Accuracy-2nd</th>
<th>Belief</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOR</td>
<td>-</td>
<td>-.50**</td>
<td>.58**</td>
<td>-.61**</td>
<td>-.61**</td>
<td>.53**</td>
<td>.32</td>
<td>.44**</td>
<td>.25</td>
</tr>
<tr>
<td>RT-1st</td>
<td>-</td>
<td>-.18</td>
<td>.22</td>
<td>.62**</td>
<td>-.12</td>
<td>-.15</td>
<td>-.03</td>
<td>-.06</td>
<td></td>
</tr>
<tr>
<td>Accept-1st</td>
<td>-</td>
<td>-.66**</td>
<td>.26</td>
<td>.96**</td>
<td>.06</td>
<td>.76**</td>
<td>.36**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Answer change</td>
<td>-</td>
<td>-.23</td>
<td>-.64**</td>
<td>-.03</td>
<td>-.61**</td>
<td>-.13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rethink time</td>
<td>-</td>
<td>-.22</td>
<td>-.43**</td>
<td>-.05</td>
<td>-.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accept-2nd</td>
<td>-</td>
<td>.04</td>
<td>.77**</td>
<td>.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy-2nd</td>
<td>-</td>
<td>-.15</td>
<td>.55**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Belief</td>
<td>-</td>
<td></td>
<td></td>
<td>-</td>
<td>.76</td>
<td>.36**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validity</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.31</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\* \(p < .05\).
\** \(p < .01\).

\(^2\) Two items were excluded from the latter analysis because there were no changes.
negative, despite the fact that reasoners spent more time rethinking problems with an initially low FOR. Moreover, the data in Table 2 suggest that the engagement of Type 2 processes does not always produce a benefit in terms of normative outcomes. Specifically, the correlation between number of changes and final accuracy was close to zero, suggesting that participants were just as likely to change from a correct answer to a wrong one as vice versa (see Shynkaruk and Thompson (2006) for a similar finding).

Moreover, the correlation between rethinking time and accuracy at Time 2 was significantly negative: The longer participants spent rethinking their answer, the less likely they were to get it right at Time 2. We speculate that reasoners engaged in a form of Type 2 thinking that produced non-normative outcomes. In support of this assertion, time spent rethinking tended to reduce the probability that a conclusion was accepted at Time 2, \( r = -.22, p = .08 \), consistent with a large body of data indicating that reasoners engage a deliberate search for counterexamples to a conclusion (e.g., De Neys et al., 2005a; Quinn & Markovits, 2002; Thompson, 2000), a strategy that leads them to reject both valid and invalid inferences.

In sum, three sets of analyses have established a clear link between the FOR that accompanies an initial response and Type 2 analysis. When the initial FOR is low, participants are more likely to change their answer and also spend more time thinking about their answer than when the initial FOR is high. Moreover, the relationship between FOR and Type 2 thinking is not an item artifact, as it is holds even when computed separately for each item.

### 2.3.7. Determinants of the FOR

In the introduction, three possibilities were suggested as determinants of the FOR, of which two were tested in the current experiment. The first, derived from the extensive literature on metamemory, was the fluency with which the initial answer was generated. Answer fluency can be defined as the speed or ease with which an answer comes to mind (Schwartz et al., 1997; Van Overschelde, 2008). Thus, under the hypothesis that a strong FOR is engendered by fluent processing, one would expect a negative relationship between the amount of time taken to produce the initial response and FOR. The correlations reported in Table 1 clearly support this prediction: The correlation between FOR judgments and the time taken to produce the initial response is large and negative, such that fluent processing was associated with strong FOR judgments. Similar conclusions were reached using participants as the units of analysis: For this, we computed each participant’s correlation between FOR and initial response time, then computed the average correlation across participants. As expected, the correlation was negative and significantly different from zero (\( M = -.32, t(26) = 8.34, p < .001 \)), indicating that fluently generated answers were associated with stronger FOR judgments.

We also found evidence to support our second hypothesis regarding the relationship between FORs and conclusion acceptance. As in previous studies, participants in the current study showed a definite

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>Believable</th>
<th>Unbelievable</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First response</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valid</td>
<td>.88 (.022)</td>
<td>.61 (.051)</td>
<td>.75</td>
</tr>
<tr>
<td>Invalid</td>
<td>.78 (.050)</td>
<td>.41 (.054)</td>
<td>.60</td>
</tr>
<tr>
<td>Mean</td>
<td>.83</td>
<td>.51</td>
<td>.67</td>
</tr>
<tr>
<td><strong>Second response</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valid</td>
<td>.87 (.029)</td>
<td>.59 (.053)</td>
<td>.73</td>
</tr>
<tr>
<td>Invalid</td>
<td>.76 (.055)</td>
<td>.33 (.044)</td>
<td>.55</td>
</tr>
<tr>
<td>Mean</td>
<td>.82</td>
<td>.46</td>
<td>.64</td>
</tr>
<tr>
<td><strong>Free response</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Valid</td>
<td>.89 (.025)</td>
<td>.62 (.051)</td>
<td>.75</td>
</tr>
<tr>
<td>Invalid</td>
<td>.69 (.054)</td>
<td>.31 (.048)</td>
<td>.50</td>
</tr>
<tr>
<td>Mean</td>
<td>.79</td>
<td>.47</td>
<td>.63</td>
</tr>
</tbody>
</table>

*Note:* standard errors are in parentheses.
acceptance bias, accepting 67% of conclusions at Time 1 and 64% at Time 2, even though it would have been logically correct to do so only 50% of the time. A possible reason for this is a stronger FOR associated with accepting, rather than rejecting an item. The correlations in Table 1 support this hypothesis: FOR is positively correlated with probability of acceptance at both Time 1 and Time 2.

Two additional determinants of the FOR are suggested from the matrix in Table 1. Each item was coded with respect to the believability and validity of its conclusions, with positive integers denoting believable and valid conclusions. As expected, both factors were positively correlated with the probability of acceptance (i.e., reasoners were more likely to accept believable than unbelievable conclusions and more likely to accept valid than invalid conclusions). Both variables also correlated with FOR: the FOR was higher for believable than unbelievable conclusions and for valid than invalid conclusions. One hypothesis is that these relationships are mediated by the relationship between FOR and conclusion acceptance, such that FOR would be higher for any variable that increased conclusion acceptance. To test this hypothesis, we re-computed the correlations between believability, validity, and FOR, partialling out conclusion acceptance; neither correlation was significant, $r(61) = .06$. In contrast, the relationship between FOR and acceptance remained reliable after partialling out belief and validity, $r(60) = .33, p = .009$. Thus, it appears that the FOR is higher for conclusions that are accepted relative to those that are not, regardless of the factors, i.e., validity and believability, that contributed to accepting the conclusion.

2.3.8. Summary: FOR, Type 2 thinking, and answer fluency

The foregoing analyses demonstrated a clear and consistent relationship between the FOR associated with an initial judgment and two measures of analytic engagement: Reasoners spent longer rethinking an answer that was produced with a low FOR and were more likely to change such an answer relative to one that was produced with a high FOR. There is also clear evidence linking the fluency associated with generating the initial answer to the FOR: Fluently generated responses produced a strong FOR. Moreover, these relationships were strong in terms of the absolute amount of variance explained and were observed using both items and participants as the units of analysis. In short, these data provide compelling support for a monitoring process that is responsive to the fluency with which initial answers are produced and that, in turn, signals the need for analytic thinking to be engaged.

2.3.9. Confirmatory analyses

The paradigm we employed to test the relationship between FOR, Type 2 thinking, and fluency is novel and represents a substantial change to the usual paradigms used to measure reasoning (i.e., generating a single answer to a problem with no time constraints). In this section, we address potential concerns about our method. The first is the degree to which it accomplished the goal of eliciting intuitive responses; the second concern is the potential for the final answer and time spent rethinking to have been altered by having made the initial response and FOR rating.

Descriptive data indicate that we were successful in eliciting participants' first responses. In the two response condition, participants indicated that they responded with the first answer that came to mind on close to 100% of trials ($M = 99.64$). Moreover, responses in this condition were produced substantially faster ($M = 7.2\,\text{s}$, $sd = 2.2$) than responses in the free time condition ($M = 10.8\,\text{s}$, $sd = 5.8$; $t(58) = 3.26, p = .001$). Thus, both in terms of self-report and behavioural data, it appears that participants complied with the instruction to give the first answer that came to mind.

The inclusion of the free-response and intuitive-response groups allowed us to rule out concerns about reactivity of measurement. First, we compared responses given in the two-response condition to those provided in the free-response condition. If providing the first response and FOR judgement altered the outcome of participants' reasoning processes we would expect to observe differences between the two groups. The data for the two groups are reported in Table 2; for the sake of completeness, the first responses given in the two-response group are also included. The analysis, however, just compares the second response to the control group. These data were analysed using a 2 (conclusion believability) \times 2 (validity) \times 2 (test condition) mixed ANOVA.

As is clear from the Table, responses in the two conditions were nearly identical both in terms of the overall rates of acceptance as well as in terms of their pattern across the belief by validity cells. As expected, the main effects of belief, validity, and their interaction were all significant, $F(1, 58) \geq 20.3$, $p \leq .001$.
More importantly, however, the rates of acceptance did not differ between the two conditions, \( F < 1 \), nor did response condition enter into any higher-order interactions, all \( F's < 1 \). Given that the final answer in the two-response condition was virtually identical to that given under free time conditions, it seems reasonable to conclude that there was little reactivity produced by giving the initial response.

Finally, we present a set of analyses to rule out the possibility that making the initial FOR judgment changed participants’ approach to the task. For example, it is possible that participants thought longer about a problem and were more likely to change their answers to that problem because they had given it a low FOR rating. To rule out this possibility, we compared rethinking times and answer changes in the two-response group to the FOR ratings given by a different set of participants, namely those in the intuitive-response condition. The unit of analysis was items: Across the 64 items, the correlations between FOR's in the intuitive-response group and rethinking times and answer changes in the two-response group were significant and negative, \( r's \leq -.27, p \leq .035 \). Finally, we also correlated FOR's in the intuitive response condition with total thinking time in the free response condition, \( r = - .30, p = .015 \). Given that these FOR's and thinking times were produced independently, reactivity cannot be an explanation for the observed relationships.

In summary, the evidence converges on the conclusion that the two-response procedure did not change participants’ responses or thought processes, namely that the pattern of final responses in the two-response group did not differ from that produced under a more traditional, single-response paradigm and that thinking times and answer changes in the two-response group were predicted by FOR's in a separate group of participants.

### 2.3.10. Final judgments of confidence (FJC)

Although they were not the primary focus of our experiment, for the sake of completeness, we provide a brief overview of reasoners’ FJC. Relative to the FOR at Time 1 (\( M = 5.30, sd = 1.27 \)), reasoners expressed more certainty at Time 2 (\( M = 5.70, sd = 1.19 \)), \( t(29) = 3.55, p = .001 \). An analysis of the FJC given in the two-response and free-response groups indicated that reasoners were more confident for believable (\( M = 5.93 \)) than unbelievable conclusions (\( M = 5.70 \)), \( F(1, 29) = 8.70, MSE = .40, p = .005 \) and for valid (\( M = 5.88 \)) than invalid conclusions (\( M = 5.75 \)), \( F(1, 29) = 7.60, MSE = .299, p = .008 \). The interaction between them was not reliable, \( F < 1 \), nor were any interactions with response condition, \( F's < 1 \).

### 2.4. Discussion

The goal of Experiment 1 was to provide an answer to the question of why, all other things being equal, some initial answers are kept with little analysis, whereas others are analysed more extensively. Consistent with the Metacognitive Dual Process Theory:

1. We established a clear relationship between a metacognitive judgment associated with an initial answer and the degree of Type 2 engagement. Although we are not the first to consider the possibility that metacognition plays a role in Type 2 thinking (Alter et al., 2007; Simmons & Nelson, 2006), we have provided the first evidence to substantiate a relationship between a measureable metacognitive construct and two measures of Type 2 thinking.

2. The Feeling of Rightness was, in turn, predicted by answer fluency; fluently generated answers produced stronger FORs than their less fluent counterparts. These data are consistent with much evidence in the metamemory literature linking retrieval fluency to metacognitive judgments such as the Feeling of Knowing or Judgements of Learning. Again, although we are not the first to speculate about a possible relationship between fluency and metacognition in reasoning (Alter et al., 2007; Simmons & Nelson, 2006), we have provided the first concrete evidence to verify that relationship.

3. We found evidence that the FOR is multiply determined, not only by fluency, but by acceptance, such that it was higher for conclusions that reasoners accepted than those they rejected. This was true even after partiailling out the variance due to conclusion validity and belief, both of which are correlated with conclusion acceptance.
In addition, our data shed light on the question of whether or not normative responding is a useful measure of Type 2 thinking (Evans, 2007b; Stanovich, 2009), given that much of the analysis engaged by our participants did not appear to be guided by the rules of logic or principles of logical necessity (despite having been instructed to do so). In many cases, the answer that participants gave after additional thought was the same as the answer given with little thought, suggesting that they were generating support for their first conjecture (Evans, 1996; Shynkaruk & Thompson, 2006; Stanovich, 2009; Wilson & Dunn, 2004). Thus, rather than trying to reformulate the problem, reasoners appeared to spend time rationalising their initial answer, which, in turn, increased confidence in that answer (Koriat, Lichtenstein, & Fischhoff, 1980; see also Gill, Swann, and Silvera (1998) and Tsai, Klayman, and Hastie (2008) for related ideas).

Moreover, when participants changed their answer, they were just as likely to change to a normatively incorrect as to a normatively correct answer, confirming that Type 2 processing does not always produce normative outcomes. As was discussed above, a possible explanation of this is that re-thinking time was associated with rejecting conclusions, which is consistent with models that propose that searching for counter-examples is one of the mechanisms mediating conditional reasoning (e.g., Cummins et al., 1991; De Neys, Schaeken, & d’Ydewalle, 2003; Grosset, Barrouillet, & Markovits, 2005; Thompson, 2000). Since this search is deliberate and WM demanding (De Neys et al., 2005a), it must be considered a Type 2 response. Thus, although it is clear that some reasoners do produce normative answers to conditional inference problems, especially those of higher reasoning capacity (De Neys et al., 2005a; Evans, Handley, & Bacon, 2009; Stanovich, 1999), it is also clear that much of the thinking engaged by reasoners on this task had a different outcome.

Finally, we were able to rule out a number of alternative explanations for our findings. First, the relationship between FOR and Type 2 engagement was not an item difficulty effect, as the relationship was observed when correlations were computed separately for each item. Second, two pieces of evidence addressed a possible concern regarding our success in eliciting participants' first responses: Participants overwhelmingly indicated that they had, indeed, responded with the first answer that came to mind, and moreover, the time required to make the initial response was just over half as long as the time taken in the free time. Third, we provided solid evidence that giving an initial response and FOR judgment did not change the manner in which the task was performed: Responses in the free time group were very similar to the final responses of the two-response group and the relationship between FORs, rethinking time, and probability of answer change held when computed across different groups of participants.

3. Experiment 2

Experiment 2 was intended as a replication and extension of Experiment 1 using different types of items and a modified conditional reasoning task. There were two primary objectives. The first was to increase the complexity of the task and allow more scope for Type 2 processes to be recruited. Thus, in contrast to Experiment 1, which employed simple conditional statements with familiar content, we shifted to a three-term format with nonsense terms used to link the premises, e.g.,

If something is a rose, then it has a gebber.
If something has a gebber, then it is a flower.
MP: Something is a rose. Therefore, it is a flower.

Manipulating the placement of the terms allowed us to create believable, unbelievable, and neutral versions of the each inference for each premise pair, e.g.,

If something is a flower, then it has a gebber.
If something has a gebber, then it is a rose.
MP: Something is a flower. Therefore, it is a rose.

This allowed us to increase the similarity of the items to each other in terms of familiarity, number of words, etc. As such, we were able to achieve our second objective, which was to reduce the possibility that the relationship between FOR and Type 2 engagement could be attributed to item artifacts
Finally, in this study, each conditional premise was presented only once, eliminating any variability in FOR judgments that may be caused by stimulus repetition. This also reduced the number of problems participants solved from 64 to 12 and thereby reduced the possibility of fatigue effects. For this reason, we did not include an intuitive-response group, as we did not have enough sufficient number of items for the analysis.

Otherwise, the design was the same as in Experiment 1. Again, it was hypothesized that the FOR accompanying an initial decision would predict Type 2 engagement as measured by the amount of time spent rethinking the initial answer as well as the probability that the initial answer would change. We once again tested the hypothesis that FOR would vary as a function of the fluency with which the initial answer was produced, as well as the probability of accepting a conclusion.

3.1. Method

3.1.1. Participants

Forty-eight participants (32 females, 16 males) took part in this experiment in exchange for research related course credit or CDN $5.00; again, the former were introductory psychology students and the latter were members of the University community who responded to advertisements. Ages ranged from 18 to 29 years (M = 21.08; SD = 3.22). Participants were randomly and equally (24 per group), assigned to one of two conditions (free- or two- response). All the participants indicated that they gave their first response on 90% or more of trials, so none were replaced.

3.1.2. Materials

As in Experiment 1, each participant completed a conditional argument task presented on a high-resolution computer monitor. For this task, participants were asked to evaluate the validity of one of the four conditional statements: Modus Ponens (MP), Modus Tollens (MT), Affirming the Consequent (AC), and Denying the Antecedent (DA), which varied in terms of belief (i.e. neutral, believable, and unbelievable), e.g.,

MP: Something is a rose. Therefore, it is a flower.
MT: Something is not a flower. Therefore, it is not a rose.
AC: Something is a flower. Therefore, it is a rose.
DA: Something is not a rose. Therefore, it is not a flower.

In the example above, the two valid inferences (MP and MT) are believable, whereas the two invalid inferences (AC and DA) were not. A second version of the problem was created to counterbalance believability and logical form by changing the order of the first and last terms, as illustrated below:

MP: Something is a flower. Therefore, it is a rose.
MT: Something is not a rose. Therefore it is not a flower.
AC: Something is a flower. Therefore, it is a rose.
DA: Something is not a flower. Therefore, it is not a rose.

In this version, the valid inferences are unbelievable and the invalid inferences were believable. Neutral versions of the problems were created by adding another non-sense word to the premises, so that the conclusions expressed a relationship between a familiar and unfamiliar term:

MP: Something is a zowel. Therefore, it is a flower.
MT: Something is not a flower. Therefore it is not a zowel.
AC: Something is a flower. Therefore, it is a zowel.
DA: Something is not a zowel. Therefore, it is not a flower.
The believability of the conclusions was established by pre-testing. We chose 12 pairs of conditional premises for this study, which could be presented in one of three belief conditions and with one of four inferences, producing a total of 144 problems. From these 144 problems, 12 different problem sets were constructed; each set included one version of each of the four inferences in each belief condition, all having different contents. In addition to these 144 problems, four practice problems were chosen from the pretest items; these used different contents than the test problems. These problems included a MP believable problem, a MT unbelievable problem, an AC neutral problem, and a DA neutral problem.

3.2. Procedure

Each participant was presented with each of the 12 premise pairs. Of these 12 problems, there were four problems in each belief condition; each of these four problems was presented with a different inference. Across participants, each pair of conditional premises appeared equally often with each belief by inference combination.

The procedure was the same as in Experiment 1 with the following exceptions: Participants evaluated 16 conditional statements: four practice and 12 test statements. Testing of each participant took approximately 10 min. The four practice problems were presented in random order and were included to ensure that participants were familiar with these types of questions before data were collected.

The instructions were also the same as those provided in Experiment 1, save those pertaining to the unique elements of this experiment. That is, participants were told that they would solve 16 problems, each consisting of two if-then statements (e.g. If something is a cigarette, then it contains ramadions; If something contains ramadions, then it is addictive) followed by a premise statement and conclusion (e.g. Something is a cigarette. Therefore, it is addictive). They were told not to be concerned if some of the terms in the problems were unfamiliar; as before, they were instructed to choose “yes” if the conclusion followed logically from the premises and “no” if it did not. The remaining instructions were identical to the instructions presented to the participants in Experiment 1.

3.3. Results

3.3.1. Scoring

The data were scored in the same fashion as Experiment 1 and the RT’s were converted to log10 prior to analysis.

3.3.2. FOR and Type 2 thinking

Here, we present the results of several sets of analyses demonstrating a link between FOR and our measures of Type 2 thinking; these analyses pertain to the two-response group only. First, as in Experiment 1, we identified trials on which participants changed their answers and compared the mean FOR and RT’s to trials on which participants did not change. On average, the participants changed about 22% of their answers (M = 2.62/12, sd = 2.43). As in Experiment 1, and consistent with our hypotheses, the FOR was lower (M = 3.55, sd = 1.33 vs. 4.71, sd = 1.14) and rethinking times were longer (M = 15.71 s, sd = 7.18 vs. 10.08 s, sd = 3.42) when participants changed answers than when they did not, t(17) = 3.88, p < .001.3 Again, the link between FOR and analytic engagement is clear, with a lower FOR and longer RT’s associated with changing responses from Time 1 to Time 2.

To test the relationship between FOR and re-thinking time (M = 9.94 s, sd = 3.61), we computed, for each participant, their correlation between FORs and rethinking times. When averaged across participants, the mean correlation was negative and significantly different from zero (Mr = −.28, t(22) = 3.81, p = .001). That is, a high FOR was once again associated with shorter rethinking times than a low FOR.

We performed a similar analysis to test the relationship between FOR and normative accuracy. For this, we identified items for which the participants gave a normatively correct answer and items on which they did not, and then compared the mean FORs for the two sets of responses. As was the case

---

3 Note that six participants did not change any answers and so were excluded from this analysis.
in Experiment 1, the relationship was small but positive, with a higher FOR associated with correct ($M = 4.95, sd = 1.30$) than incorrect ($M = 4.75, sd = 1.47$) responses, $t(23) = 1.81, p = .041$, one-tailed.

3.3.3. FOR, fluency, and probability of acceptance

We performed similar analyses to test the relationship between the FOR, fluency (i.e., the amount of time required to produce the initial answer) and the probability of accepting a conclusion. First, for each participant, we computed a correlation between FOR and the time required to generate the initial response. When averaged across participants, the mean correlation was negative and significantly different from zero ($Mr = -.30, t(22) = 5.67, p < .001$), consistent with the hypothesis that FORs are mediated by the fluency of producing the initial answer.

As was the case in Experiment 1, participants tended to accept more conclusions than they rejected both at Time 1 ($M = .65, sd = .24$) and Time 2 ($M = .65, sd = .25$). To test the relationship between FOR and the probability of accepting an inference, we calculated the mean FOR for those trials on which participants accepted vs. rejected an inference at Time 1. As was the case in Experiment 1, we once again observed that the FOR was higher on trials where participants accepted conclusions ($M = 4.82, sd = 1.36$) than when they rejected conclusions ($M = 4.30, sd = 1.61$), $t(21)^4 = 2.50, p = .01$, one-tailed.

3.3.4. Confirmatory analyses

As was the case in Experiment 1, the descriptive data indicated success in eliciting participants’ first responses. In the two response condition, participants indicated that they responded with the first answer that came to mind on close to 100% of trials ($M = 99.58$). Moreover, the time taken to produce the first response in this condition ($M = 10.2 s, sd = 3.24$) was about half as long as the time taken to produce a response in the free time condition ($M = 18.52 s, sd = 8.16$), $t(46) = 5.64, p < .001$). Again, both in terms of self-report and behavioural data, it appears that participants complied with the instruction to give the first answer that came to mind.

To verify that the two response procedure did not change participants’ reasoning, we compared the final responses given in the two-response group to the responses given in the free-time control group. These data are presented in Table 3. Again, for the sake of completeness, the first responses given in the two-response group are also included.

The analysis was a $3 \times 2 \times 2$ mixed ANOVA. As is clear from the Table, responses in the two conditions were similar both in terms of the overall rates of acceptance as well as in terms of their pattern across the belief by validity cells: The rates of

---

4 Two participants were excluded because they did not reject any conclusions.
acceptance did not differ between the two conditions, \( F < 1 \), nor did response condition enter into any higher-order interactions, \( F_{s} < 2.10, p > .13 \). Given that the final answer in the two-response condition was virtually identical to that given under free time conditions, it seems reasonable to conclude that there was little reactivity produced by giving the initial response.

### 3.3.5. Final judgments of confidence

Finally, we present a brief summary of reasoners’ FJC’s. As was the case in Experiment 1, reasoners expressed more certainty in their FJC (\( M = 5.69, sd = 1.19 \)) than in their FOR (\( M = 4.88, sd = 1.56 \)), \( t(23) = 5.20, p < .001 \). An examination of the FJC produced in the free- and two-response conditions produced a main effect of believability, \( F(2,92) = 8.82, MSE = .714, p < .001 \): confidence was higher for believable (\( M = 5.90 \)) than unbelievable (\( M = 5.53 \)) or neutral conclusions (\( M = 5.46 \), \( t(47) > 3.28, p < .002 \), which did not differ, \( t < 1 \). No other main effects or interactions were reliable, \( F(1,23) < 2.14, p > .12 \).

### 3.4. Conclusions

In this experiment, we constructed the items to be as similar as possible to each other to eliminate extraneous item characteristics as a potential explanation for the relationship between the FOR and Type 2 thinking. We also used a more challenging task to allow a broader scope for the observation of Type 2 processes. Neither of these changes modified our earlier conclusions. Consistent with our Metacognitive Reasoning Theory, a clear and consistent relationship between the FOR associated with an initial judgment and two measures of analytic engagement has emerged: Answers produced with a low FOR were subject to longer analysis and were more likely to change relative to ones produced with a high FOR. In turn, there was clear evidence FOR to both the fluency associated with generating the initial answer to the FOR and the probability that the conclusion was accepted. In short, these data provide compelling support for a monitoring process that is responsive to the fluency with which initial answers are produced and that, in turn, signals the need for analytic thinking to be engaged.

### 4. Experiment 3

The preceding two Experiments provided evidence to indicate that answer fluency mediates FOR, such that the more fluently an answer can be retrieved, the stronger the FOR that accompanies it. The procedure used allowed us to compare, within a block of trials, the consequences of relative fluency for FOR, and in turn, how the FOR impacts Type 2 thinking. An alternative test of the hypothesis would be to manipulate characteristics of the answers believed to affect FOR judgments and to then compare signatures of Type 2 thinking across conditions. This was the goal for the next two experiments.

The current experiment tested the hypothesis about the contribution of conflicting information for FOR judgments. The task used was adapted from De Neys and Glumicic (2008). This is a variant of the classic paradigm used to study base-rate neglect (Kahneman & Tversky, 1973) in which participants are presented with two pieces of information, namely, the prior probability (base rate) that an individual belongs to one of two categories and a personality description of a particular individual. Participants are asked to estimate, based on these two sources of information, the probability that the individual described is a member of one of the categories, e.g.,

In a study 1000 people were tested. Among the participants, there were 3 nurses and 997 doctors. Paul is a randomly chosen participant of this study. Paul is 34 years old. He lives in a beautiful home in a posh suburb. He is well spoken and very interested in politics. He invests a lot of time in his career.

What is the probability that Paul is a doctor?

The two sources of information can be congruent (i.e., the description is consistent with the largest category), incongruent (i.e., the description is consistent with the smaller category) or neutral (i.e., the personality description does not favour one category or the other).
This paradigm thus allowed a straightforward test of the hypothesis that the FOR is sensitive to conflict between two inputs. This latter prediction is derived from Dual Process Theories that posit a special status to problems in which answers derived from Type 1 and Type 2 processes conflict (e.g., Amsel et al., 2008; Ball et al., 2006; De Neys & Glumicic, 2008; Evans, 2007a; Sloman, 1996), as when the base rate and personality description cue different responses. These theories assume that Type 1 and Type 2 processes are engaged from the outset, so that two conflicting answers may be produced (Sloman, 2002). In order to produce a conflict with fast, automatic Type 1 processes, this initial Type 2 processing needs to be relatively low-level and shallow and there needs to be a set of processes in place to detect such conflicts when they occur (De Neys & Glumicic, 2008). On this view, we propose the FOR as a means of monitoring this output, such that conflicting outputs lower FORs. Thus, FORs should be lower for incongruent than for congruent items. Once detected, the need to resolve the conflict is posited to give additional Type 2 processing; this additional processing is not engaged in the absence of conflict (Evans, 2007a), so that rethinking times should be longer for incongruent than congruent items.

A second goal of this experiment was to replicate our findings from Experiments 1 and 2 using a task that was as different as possible from the deductive reasoning paradigms used there. In Experiments 1 and 2, participants were required to apply a pre-specified criteria (logical validity) to their answers, they evaluated presented conclusions as opposed to generating them, and they were asked for a binary as opposed to a continuous response. In Experiment 3, we changed all of these features in order to generalize the relevance of FOR beyond deductive reasoning. In addition, because participants generated a continuous, rather than a dichotomous response, the current experiment allowed us to explore the relationship between the FOR and the degree, as well as the probability, of answer change.

Otherwise, this experiment used a procedure similar to the previous ones. Participants made two judgments: one quickly and the other after reflection. For this study, we gave participants a deadline for making their first response as an added incentive to provide the first answer that came to mind. They provided FOR ratings immediately after their first judgment and this FOR was used to predict the degree of answer change as well as the amount of re-thinking time. In addition, we included a control group that were allowed to respond in free time conditions.

4.1. Method

4.1.1. Participants

One hundred and twenty-eight participants with a mean age of 22 years (sd = 5) were tested; 64% (82) were female. Participants were randomly assigned to either the two response group (N = 64) or free response group (N = 64). Three participants in the two-response condition were replaced because they indicated that they had not provided their initial answer on 5 or more trials. Equal numbers in the two groups either received partial course credit or were paid CAN $5.00.

4.1.2. Materials

Two versions of 18 base rate problems were presented on a high-resolution computer monitor using E-Prime. As illustrated above, each problem described a sample taken in a study, the probability (base rate) of belonging to that sample, as well as a personality description of a specific individual who was part of the study. The problems were adapted from De Neys and Glumicic (2008) with the following modifications: (1) A few of the individuals’ names were changed to make them more gender neutral; (2) Two versions of each problem were created by switching the large and small base rate numbers so that the same personality description could be presented in both the congruent and incongruent conditions; and (3) Instead of asking participants to make a binary choice of two categories, participants were asked to estimate the probability that the individual belonged to one of the categories.

4.1.3. Procedure

Problems appeared on the monitor one at a time followed by a text box in which participants provided their probability estimate (out of 100). After their response was made, a likert-scale appeared. At Time 1, the scale measured how “right” the participants felt about their answer, where 1 corresponded
to doesn’t feel right at all and 7 corresponded to feels very right (i.e., FOR). At Time 2, participants rated their confidence in their final answer on a scale where 1 corresponded to not at all confident and 7 corresponded to extremely confident (i.e., FJC).

The problems were presented in a different random order for each participant. There were three base rate probabilities, as per De Neys and Glumicic (2008); within each congruency condition, two problems were presented with a 995/5, two with a 996/4, and two with a 997/3 base rate ratio. For each ratio, participants were asked to make one judgment about the smaller category and one about the larger category. This was counterbalanced across participants, so that the larger and smaller version of each problem was tested equally often, allowing the congruent and incongruent scenarios to be equated across participants.

Participants were tested individually. The instructions were similar to those used in the first two studies save for those parts that described unique elements of the current experiment; these were adapted from De Neys and Glumicic (2008). Specifically, prior to beginning the study, participants were told:

In a big research project, a number of studies were carried out where short personality descriptions of the participants were made. In every study there were participants from two population groups (e.g., carpenters and policemen). In each study, one participant was drawn at random from the sample. You’ll get to see the personality description of this randomly chosen participant. You’ll also get information about the composition of the population groups tested in the study of question.

For each problem, you’ll be asked to provide us with a probability judgment, out of 100, indicating what you believe the likelihood is that the randomly chosen participant belongs to the specified group. For example, if you are sure that the participant belongs to the specified group, you should choose a number closer to 100. Conversely, if you are sure that the participant does not belong to the specified group you should choose a number closer to 0. Finally, if you are unsure about which group the participant belongs to, then you should choose a number close to 50. For all problems, the group that you are asked to give the probability for was randomly chosen out of the two groups presented.

In the one-response condition, participants gave a single response without time constraint. In the two-response condition, participants were told to provide two answers; again, the first answer was to be their first inclination or instinct. It was stressed that they were to provide us with the first answer that came to mind and were given a 12 s deadline to respond.5 To emphasise the need to respond quickly, the problem was italicized and changed colour after 12 s elapsed, and participants were asked to enter their answer immediately. After entering their answer, they provided a Feeling of Rightness estimate using the scale described above. They were then asked whether they had, indeed, responded with their first answer. Finally, they were allowed all the time they needed to make their final answer and final confidence judgment. The next trial appeared when they pressed the space bar.

Prior to testing, participants were presented with a neutral practice problem in order to familiarize them with the procedure. After the practice problem, it was emphasized to participants in the two response condition that they should not second-guess their first answer. All participants were allowed to ask any questions and proceed to the next trial by pressing the space bar. Testing took approximately 25 min.

4.2. Results

4.2.1. Scoring

To facilitate interpretation, the scores for the low-base rate items were subtracted from 100 so that high scores always indicated estimates that were close to the base rate and low numbers reflected estimates that deviated from the base rate. It is not possible to derive a measure of normative accuracy for these data, as that would require combining the base rate with probability estimates for the

---

5 It should be noted that the 12-s deadline was chosen on the basis of pilot studies. In these studies, we observed that 15 s was too long and did not create enough urgency and that 10 s was too short because it did not allow enough time to read the problem.
personality descriptions, which were not available. All RT measures were converted to $\log_{10}$ prior to analysis. In the two-response condition, responses for which participants indicated that they did not reply with their first intuition (3.6%) were eliminated from the analysis.

4.2.2. Analysis strategy

Once again, the data are reported in several sections. The first data reported are those that test the hypotheses about the relationship between FOR, Type 2 engagement, and fluency. We then present data to test the hypothesis about the special status of conflict items. Each of the subsequent analyses diffuses potential counter-arguments to the conclusions reached on the basis of these initial findings, including potential arguments about our success in eliciting participants’ intuitive answers, potential reactivity of the first judgment on the final one, etc. We end with a brief summary of reasoners’ FJC’s.

4.2.3. Relationship between FOR and two measures of Type 2 engagement

In this section, we tested the hypothesis that initial responses accompanied by a low FOR would be associated with more Type 2 thinking than responses associated with a high FOR. To do this, we examined the relationship between the FOR at Time 1, the amount of time spent re-thinking a problem, and the probability and degree of answer change from Time 1 to Time 2.

As in our earlier experiments, we began with a participant-based strategy. Again, we identified trials on which participants changed their answers and compared the mean FOR and RT’s to trials on which participants did not change. Unlike our previous studies, participants changed the majority (61%) of their answers. More importantly, consistent with our hypotheses, the FOR was lower when participants changed answers ($M = 4.55, sd = 1.04$) than when they did not ($M = 5.09, sd = 1.16$), $t(58) = 4.67, p < .001$, and as illustrated in Fig. 1, this was true for all three problem types, $t > 2.45, p < .012$. In addition, because the response data were continuous, rather than dichotomous, we were also able to compute the FOR as a function of degree of change; for this, we determined the median degree of change for all trials on which participants changed their answers ($M = 25.3$). The FOR for changes greater than the median ($M = 4.04, sd = 1.15$) was lower than the FOR for small changes ($M = 4.62, sd = 1.01$), $t(56) = 4.29, p < .001$, which, in turn, was lower than the FOR where no change was made at all, $t(51) = 2.19, p = .033$. Across all participants, the mean correlation between degree of change and FOR was $-.22, t(64) = 5.84, p < .001$.

Fig. 1. FOR for changed and unchanged answers as a function of problem congruency.

6 Again, not all participants changed an answer, so that 5 were not included in this analysis.

7 Once again, participants who did not contribute observations to both cells of the comparison were excluded from the analysis.
We did a similar analysis to test the relationship between FOR and re-thinking time. We computed, for each participant, their correlation between FOR and rethinking times. When averaged across participants, the mean correlation was negative and significantly different from zero ($M_r = -.18, t(63) = 5.05, p < .001$). That is, a high FOR was once again associated with shorter rethinking times than a low FOR.

We then replicated these findings using an item-based strategy. Recall that each personality description was presented as a congruent and incongruent item to different groups of participants; this meant that there were two sets of 18 items, each of which was presented to 32 participants. For each of these 36 items, we computed the mean FOR at Time 1 ($M = 4.74, sd = .42$), the mean time spent rethinking ($M = 16.31$ s, $sd = 2.55$), and the absolute degree of change between Time 1 and Time 2 ($M = 15.92, sd = 4.44$). The correlation matrix is presented in Table 4.

Replicating our earlier findings, there were large, negative correlations between FOR measured at Time 1, the amount of time spent rethinking, and the degree to which an answer changed during the rethinking period. That is, participants spent longer thinking about and made larger adjustments to items that engendered a low FOR at Time 1 than those for which the FOR was high. These data are consistent with a model in which an initial answer to a problem is accompanied by a metacognitive experience, which, in turn, predicts the probability that Type 2 processes are engaged.

What was the outcome of this Type 2 engagement? Overall, reasoners gave higher probability estimates at Time 2 ($M = 71.91, sd = 19.39$) than at Time 1 ($M = 67.56, sd = 19.39$), $t(35) = 3.42, p = .002$, indicating a tendency for reasoners to shift their answers in the direction of the base rate over time. This difference ($M \approx 4.5$), however, is small relative to the absolute degree of change reported above ($M \approx 16$), suggesting that a substantial proportion of answers shifted away from the base rate at Time 1. Moreover, as is clear by the correlation between answer change and rethinking time (Table 4), the degree of change was linked to rethinking time, such that longer rethinking times produced larger changes. Thus, there is evidence that time spent rethinking an answer produces changes to that answer, but as in our previous experiments, these are not necessarily either systematic or mediated by normative standards.

These analyses provide a strong replication of the first two experiments. A clear and consistent relationship between the FOR associated with an initial judgment and two measures of analytic engagement has emerged: Answers produced with a low FOR are subject to longer analysis and are more likely to change relative to ones produced with a high FOR. In this experiment, we also established that FOR is linked to the degree as well as to the probability of change, such that a low FOR predicted a larger degree of change than a high FOR.

4.2.4. Answer fluency and the FOR

The relationship between the FOR and fluency was as expected from the first two experiments. As is clear from Table 4, there is a significant, negative correlation between the time taken to produce the initial answer and ratings of FOR, consistent with the hypothesis that fluently generated answers produced a strong FOR. A participant-based analysis replicated these findings. For each participant, we computed a correlation between FOR and the time required to generate the initial response. When averaged across participants, the mean correlation was negative and significantly different from zero ($M_r = -.22, t(63) = 6.41, p < .001$). These data provide compelling support for a monitoring process that

<table>
<thead>
<tr>
<th></th>
<th>FOR</th>
<th>RT-1st</th>
<th>Answer change</th>
<th>Rethink time</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOR</td>
<td>–</td>
<td>–.32</td>
<td>–.38*</td>
<td>–.52**</td>
</tr>
<tr>
<td>RT-1st</td>
<td>–</td>
<td>–</td>
<td>.24</td>
<td>.52**</td>
</tr>
<tr>
<td>Answer change</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.43**</td>
</tr>
<tr>
<td>Rethink time</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

* $p < .05$.
** $p < .01$, one-tailed.
is responsive to the fluency with which initial answers are produced and that, in turn, signals the need for analytic thinking to be engaged.

4.2.5. FOR and conflicting information

To test the hypothesis that FOR would be lower on conflict than non-conflict items, we compared performance on incongruent to congruent problems. Congruent and incongruent items used identical personality descriptions (counterbalanced across participants), whereas neutral items were based on different descriptions (cf. DeNys et al.). Consequently, comparisons with the neutral items are prohibited by potential item effects, and our analyses focussed specifically on the theoretically salient contrast between congruent and incongruent items.

The data are summarised in Table 5. Consistent with the hypothesis that metacognitive monitoring is sensitive to conflict (De Neys & Glumicic, 2008), FOR judgments were lower for the incongruent than the congruent problems; congruent items were also responded to more fluently at Time 1 than incongruent items. Also consistent with De Neys & Glumicic’s (2008) findings that conflict prompted analytic thinking, we found that congruent items received less analytic engagement, as evidenced by shorter rethinking times and a smaller degree of answer change.

4.2.6. Confirmatory analyses

As was the case in the first two experiments, the descriptive data indicated success in eliciting participants’ first responses. In the two response condition, participants indicated that they responded with the first answer that came to mind on close to 100% of trials ($M = 96.3\%$). Moreover, the time taken to produce the first response in this condition ($M = 13.95\text{s}$, $sd = 3.03$) was about half the time taken to produce a response in the free time condition ($M = 25.50\text{s}$, $sd = 6.36$), $t(126) = 13.6$, $p < .001$. Again, both in terms of self-report and behavioural data, it appears that participants complied with the instruction to give the first answer that came to mind.

To allay possible concerns about the reactivity of measurement, we compared responses produced in the two-response condition to those produced under more traditional circumstances. To do this, we compared the final responses given by participants in the two-response condition to the responses generated in the free response condition. These data are presented in the bottom panel of Table 6 and were analysed using a $3 \times 2$ mixed ANOVA. As is clear from the Table, responses in the two conditions were nearly identical both in terms of the overall levels of estimates as well as in terms of their pattern across the congruency conditions. Probability estimates did not differ between the two conditions, nor did response condition enter into any higher-order interactions, $Fs < 1$. Given that the final answer in the two-response condition was virtually identical to that given under free time conditions, it seems reasonable to conclude that there was little reactivity produced by giving the initial response.

Table 5

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Congruent</th>
<th>Incongruent</th>
<th>$t(63)$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOR</td>
<td>5.13 (.14)</td>
<td>4.71 (.15)</td>
<td>4.19</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>RT1 (fluency)</td>
<td>13.46 s (.39)</td>
<td>14.05 s (.41)</td>
<td>3.18</td>
<td>.002</td>
</tr>
<tr>
<td>RT2 (rethinking)</td>
<td>13.61 s (.80)</td>
<td>17.89 s (1.1)</td>
<td>5.25</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Degree of change</td>
<td>12.33 (1.64)</td>
<td>19.67 (2.01)</td>
<td>3.51</td>
<td>.001</td>
</tr>
</tbody>
</table>

*Note:* standard errors are in parentheses.

Table 6

<table>
<thead>
<tr>
<th></th>
<th>Congruent</th>
<th>Incongruent</th>
<th>Neutral</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>First response</td>
<td>83.72 (1.66)</td>
<td>47.75 (3.38)</td>
<td>72.26 (2.44)</td>
<td>67.91</td>
</tr>
<tr>
<td>Second response</td>
<td>90.19 (1.50)</td>
<td>48.44 (3.88)</td>
<td>79.88 (2.57)</td>
<td>72.84</td>
</tr>
<tr>
<td>Free response</td>
<td>88.16 (1.52)</td>
<td>47.05 (3.73)</td>
<td>76.37 (2.50)</td>
<td>71.05</td>
</tr>
</tbody>
</table>

*Note:* standard errors are in parentheses.
4.2.7. Final judgments of confidence

Once again, reasoners expressed more certainty in their FJC ($M = 5.60$, $sd = .81$) than in their FOR ($M = 4.72$, $sd = 1.02$), $t(63) = 9.57$, $p < .001$. An analysis of the FJC provided in the two-response and free-response conditions indicated a main effect of congruency, $F(2, 252) = 31.14$, $MSE = .333$, $p < .001$, such that confidence was higher for congruent ($M = 5.90$) than incongruent ($M = 5.44$) items, and lowest for neutral ($M = 5.27$) items, $t(127) \geq 2.11$, $p \leq .04$.

4.3. Conclusions

These data provide an important replication of our first two experiments and demonstrate conclusively that the relationship between the FOR and Type 2 engagement is not an artifact of the specific task parameters employed in those experiments. Whereas reasoners in the first experiments evaluated conclusions, in Experiment 3, they generated conclusions. Whereas responses in the first studies were dichotomous, they were continuous in the current study. Whereas participants were provided with strict instructions about the criteria used to evaluate conclusions in the first studies, that was not the case here. Whereas the previous participants evaluated validity, here, they estimated probability. We also note that whereas a minority of responses were changed between the first and second response in Experiments 1 and 2, in the current experiment, the majority of answers changed.

Despite these differences, the relationship between FOR and Type 2 thinking remained the same. Consistent with the Metacognitive Reasoning Theory, initial answers that engendered a weak FOR were given more Type 2 analysis than answers generated with a strong FOR. This relationship held regardless of whether Type 2 thinking was indexed by the amount of time spent thinking, the probability of changing the initial answer, or the degree to which the initial answer was changed. In turn, the FOR was mediated by answer fluency, such that a strong FOR was associated with answers that were produced relatively quickly.

We also found evidence that conflict is a determinant of the FOR. Problems in which the personality description and the base rates pointed to the same answer (congruent) were clearly advantaged both in terms of FORs and rethinking times relative to situations where the two conflicted (incongruent). De Neys and Glumicic (2008) proposed that conflict per se is enough to trigger deeper analytic processing for incongruent relative to congruent items. However, we observed that the congruent items were processed more fluently than their incongruent counterparts. Thus, it is possible that fluency mediates the relationship between conflict and Type 2 thinking, although the current data do not discriminate the two hypotheses.

De Neys and Glumicic (2008) suggested that a type of “shallow” analytic processing of the base rates in order to explain why a supposedly slow, Type 2 process can be completed in time to conflict with a supposedly faster, Type 1 process. In that regard, we note that the effects of conflict on FOR judgments does not necessarily imply a conflict between Type 1 and Type 2 responses. The incongruent problems present reasoners with two types of information: a personality description and a pair of numbers. Whilst one might be tempted to conclude that evaluating the latter requires Type 2 processes, there is no reason to suppose that comprehending the numbers requires elaborate analysis (see also Gigerenzer & Regier, 1996). That is, participants would effortlessly comprehend that a three-digit number (e.g., 995) represents a much larger quantity than a one-digit number (e.g., 5), without necessarily estimating or calculating a precise ratio.

Finally, as was the case with our first two experiments, the data from Experiment 3 suggest caution in using normative measures as indices of Type 2 thinking (Evans, 2007b, 2009; Stanovich, 2009). One might be tempted to conclude that answers close to the base rate are more “correct” than answers deviating from the base rate (De Neys & Glumicic, 2008), although technically speaking, deriving the correct answer requires combining probabilities derived from both the base rate and the personality description. Nonetheless, when the personality descriptions and the base rates provided conflicting cues, reasoners often changed their initial response in the direction of the personality description rather than in the direction of the base rate, suggesting that both “correct” and “incorrect” answers can be arrived at via deliberate analysis. Thus, the fact that a non-normative answer was given does not necessarily imply the absence of Type 2 engagement.
5. Experiment 4

Whereas in the previous study, we sought to manipulate FOR judgments by manipulating the congruence of base rates and stereotypes, in the current study we did so by manipulating the fluency with which the initial answer could be produced. Participants in this experiment were asked to solve quantified syllogisms by indicating whether or not the conclusion provided followed validly from the premises, e.g.,

Some of the nurses are magicians.
All of the winemakers are nurses.
Therefore, some of the magicians are winemakers.

Half of the conclusions were consistent with the Min Heuristic (Chater & Oaksford, 1999). The min heuristic is a non-logical strategy for evaluating conclusions based on the relative informativeness of the conclusions and premises. To be consistent with the min heuristic, the quantifier of the conclusion should be the same as the least informative premise, with “All” more informative than “Some”, which is more informative than “None”, followed by “Some not”. This strategy will produce the most informative conclusion that are consistent with a pair of premises.

As an example, the conclusion above is consistent with the min heuristic, because the quantifier of the conclusion, “some” is the same as the least informative premise (Some of the nurses are magicians). In contrast, the conclusion below violates the min heuristic, because the conclusion is less informative than either of the premises. Logically, both conclusions have the same status, in that they are consistent with, but not necessitated by, the premises.

None of the nurses are magicians.
Some of the winemakers are nurses.
Therefore, some of the magicians are not winemakers.

Chater and Oaksford (1999) documented that reasoners’ performance on syllogistic tasks is consistent with the application of the min heuristic. On the basis of their findings, therefore, we predicted that conclusions consistent with the min heuristic should be accepted more often that those that violate the heuristic, regardless of logical validity. Moreover, because the min heuristic is a “fast and frugal” strategy, conclusions that are consistent with the heuristic should be processed more fluently than those that are not, and should consequently give rise to stronger FOR judgments. As a result, min conclusions should receive less Type 2 analysis than their non-min counterparts.

For this study, we did not include a control group that provided only single responses, given that we have already provided ample demonstration that reasoners’ final judgments do not change as a function of providing the first judgment.

5.1. Method

5.1.1. Participants

Sixty-four University of Saskatchewan students participated; 56 received partial course credit for their introductory psychology course and the remainder were paid a small sum (CAN$5). Five participants were replaced for indicating that they had not responded with the first answer that came to mind on more than 10% of the trials. Sixty-three percent were female; the mean age was 20 years (sd = 3).

5.1.2. Materials

Each participant completed a syllogistic reasoning task presented on a high-resolution computer monitor using the program E-Prime (Schneider et al., 2002). Each reasoning task consisted of 17 three-term quantified syllogisms; the terms consisted of two un-repeated categories (conventionally referred to as A and C) and a category that was repeated in both premises (B). The conclusions linked
the two unrepeated terms. The first syllogism was a practice syllogism to familiarize participants with the task.

Half of the remaining 16 syllogisms had conclusions consistent with the min heuristic and half did not. Within each conclusion type, we controlled for the logical status of the conclusion, such that half of the conclusions were logically valid and half were not. We also controlled for figural effects (the order in which the three terms appear in the premises), such that there were two problems (one valid and one invalid) from each of the four Figures for each conclusion type.

For Figures 1 and 4, the terms in the conclusion were presented in the CA order; conclusions to Figure 2 and 3 premises were in the AC order. For the asymmetric problems (Figures 1 and 2), all conclusions that followed the min-heuristic were “some” conclusions and all the conclusions that did not follow the min-heuristic were “some not”. For the symmetric problems all conclusions contained the “some not” quantifier. Valid and invalid problems were matched as closely as possible on the probability that the conclusion would be accepted (based on Evans, Handley, Harper, & Johnson-Laird, 1999). To achieve these constraints, it was not possible to equate all of the valid problems in terms of the number of models or representations consistent with the premises (Johnson-Laird, 1983); to ensure that this did not jeopardize the interpretation of our findings, the analyses will also be conducted with only the invalid problems, which, by definition, allow multiple representations.

Content was assigned to the test problems in the following manner: Each A, B, and C term was filled by professions or hobbies. Each participant was presented with 16 unique combinations of professions and hobbies, with none of the categories repeated within a participant. Across participants, each combination of professions and hobbies appeared equally often for problems in each logical status and informativeness level.

5.1.3. Procedure

Participants were tested individually. Testing took approximately 20 to 25 min. Problems were shown one at a time and the premises and conclusion appeared simultaneously, with two response options, “yes” and “no” beneath the conclusion. The 16 test syllogisms were presented in a different random order for each participant. Before beginning, written instructions appeared on the computer screen. These explained the nature of the reasoning task and explained how to respond on the keyboard. With only one minor modification, the instructions and procedure was identical to that outlined in Experiment 1. Specifically, participants were reminded to give their first response prior to each problem with a “Quick Answer” cue that flashed prior to the problem being presented.

5.2. Results

On average, participants reported that they had given their first answer on 98.5% of trials; all participants reported giving their first answer on at least 14 of 16 trials. As before, all RT’s were converted to log10 prior to analysis.

5.2.1. FOR, fluency, and measures of Type 2 thinking

We began by establishing that the current data replicate the relationships noted previously between FOR, fluency, and Type 2 thinking. To test the relationship between FOR and fluency, we computed, for each participant, a correlation coefficient between FOR judgment and time to make the first response and then averaged the correlations across participants; the mean correlation ($M_r = -.19$) was significantly different from zero, $t(63) = 4.95, p < .001$. A similar computation confirmed that the correlation between FOR and rethinking times ($M_r = -.26$) was likewise reliable, $t(63) = 6.80, p < .001$. Also as in previous analyses, FOR’s were lower for answers that were changed ($M = 3.32$) than answers that were not changed ($M = 4.04$), $t(55) = 6.1, p < .001$. Thus, as expected, fluently generated answers were given higher FOR’s than less fluent answers; moreover, low FOR’s were associated with more analytic thinking, as measured by rethinking time and the probability of changing answers.

5.2.2. Min vs. non-min

Table 7 summarizes the responses given for min and non-min conclusions. As predicted, reasoners were more likely to accept the min conclusions relative to the non-min conclusions. More importantly,
the min-conclusions were processed more fluently than the non-min conclusions at Time 1. Consistent with this difference in fluency, the FOR was stronger for min than non-min conclusions. Finally, two indices of Type 2 thinking were reduced for min relative to non-min conclusions: Rethinking time and the probability of answer change were both lower for min than no-min conclusions. As was the case in our previous studies, this increase in Type 2 analysis did not produce an increase in normative accuracy between the min and the non-min problems.

Two additional sets of analyses were performed to rule out alternative interpretations of these findings. First, as indicated in the Method section, whereas all of the invalid problems afforded multiple representations of the premises, this was not the case for the valid problems. To ensure that representational complexity did not affect the outcome of the min analysis, we repeated the analysis using only the invalid problems. All of the tests produced identical outcomes except the final one involving accuracy: for the invalid problems, reasoners were more accurate for the Non-min problems (.42 vs. .15, t(63) = 6.10, p < .001); this is almost certainly due to the fact that they were more likely to incorrectly accept the min-conclusions, as indicated in Table 6. Consistent with this conclusion, for the valid problems, where accepting the conclusion produces correct answers, reasoners were more accurate with the min (.91) than the non-min (.64) conclusions, t(63) = 6.64, p < .001.

Finally, because the Min conclusions were accepted more often than their non-min counterparts, and because rate of acceptance affects FOR judgments (Experiments 1 and 2), we repeated the analyses looking just at conclusions that were accepted. Once again, the differences between the min and non-min conclusions were observed in all cases (p ≤ .055). Thus, the differences in FOR and Type 2 thinking observed between the min and the non-min conclusions cannot be attributed to differences in overall acceptance rates for the two conclusion types.

5.3. Conclusions

In this study, we manipulated a variable that we predicted would affect the fluency with which a conclusion could be evaluated and found that when asked to give an intuitive response, reasoners were faster to evaluate the min relative to the non-min conclusions. Consistent with this increase in fluency, the FOR was higher for min than non-min conclusions, and moreover, signatures of Type 2 thinking were less pronounced for the min than non-min conclusions. These data provide further evidence for a monitoring processes that is based on answer fluency, namely that fluent responses give rise to a strong FOR, which in turn, signals that extensive re-analysis is not necessary.

6. General discussion

Until recently, the issue of monitoring and control processes in reasoning has received little attention. Indeed, the need to include such a mechanism into models of reasoning has only recently been acknowledged (De Neys & Glumicic, 2008; Evans, 2009; Thompson, 2009, 2010). In the current paper, we adapted a paradigm from the metamemory literature (Koriat & Goldsmith's the quantity-accuracy profile) that allowed us to undertake a detailed analysis of monitoring in the context of reasoning. As is the case with the QAP, our two-response paradigm required participants to generate two responses and to provide a measure of confidence in the initial response that was used to predict subsequent
performance. In four experiments, we found that the initial confidence judgment, called the FOR, predicted the length of time that reasoners spent reaching a final answer, as well as the probability that they changed their answer in the interval.

We have argued that a monitoring mechanism, such as described above, is a necessary addition to Dual Process Theories of reasoning (e.g., Evans, 2006; Kahneman, 2003; Sloman, 2002; Stanovich, 2004). Specifically, some type of monitoring mechanism is required to explain why, all other things being equal, some problems receive more Type 2 analysis than others. According to the Metacognitive Reasoning Theory, the initial Type 1 output is accompanied by a metacognitive judgment called the FOR, which, in turn, determines the extent of analytic engagement (as indexed by rethinking time) as well as the outcome of that engagement (as measured by the probability and extent of answer changes). When the FOR is compelling, the initial answer is unlikely to change; when the FOR is less strong, the initial answer is more likely to change. Thus, after equating for other variables known to predict Type 2 engagement, such as cognitive capacity (De Neys, 2006a; Stanovich, 1999), instructions (e.g., Daniel & Klaczynski, 2006; Evans et al., 1994; Vadeboncoeur & Markovits, 1999), and epistemic thinking dispositions (Klaczynski & Robinson, 2000; Stanovich, 1999, 2009), the probability that a reasoner on a given occasion reconsiders their initial answer is determined by the FOR that accompanies that answer.

By extension, we would argue that a monitoring mechanism, such as the one described in this paper, will need to be incorporated into single-process theories of reasoning as well. These theories differ greatly from each other, although most emphasize the adaptiveness of heuristic responses (e.g., Gigerenzer & Brighton, 2009; Oaksford & Chater, 2007). Unlike Dual Process Theories, these perspectives do not make a strong distinction between analytic (Type 2) and heuristic (Type 1) processes and vary in terms of the extent to which they assume that heuristic processes arise automatically or deliberately (Frederick, 2002; Gigerenzer, 2008; Oaksford & Chater, 2007; Osman, 2004). Nevertheless, it is clear that only some participants give the heuristic response to a problem (Newell, Weston, & Shanks, 2003), and that there is considerable variability in the extent to which participants’ final answers reflect deliberate processing (Lee & Cummins, 2004). Thus, regardless of whether it is assumed that Type 2 processing differs qualitatively from Type 1 processing, it seems reasonable to posit a monitoring mechanism that signals the need for further processing.

6.1. Types 1 and 2 answers

In our experiments, and unlike the QAP, the first response was to be given quickly, intuitively, and with a minimum of thought. This was done to maximize the probability that participants provided the first answer that came to mind. According to DPT, such an answer would likely have been produced by Type 1 processes. It is important to note, however, that whilst we have good evidence that participants provided us with their first response to the reasoning problems, we did not provide evidence that those answers were produced exclusively by Type 1 processes as assumed by Dual Process Theories. For that, we refer to others, who have demonstrated that signatures of analytic, Type 2 processes increase with the amount of time available and that the speeded response paradigm elicits responses that favour heuristic over analytic processes (De Neys, 2006b; Evans & Curtis-Holmes, 2005; Finucane et al., 2000; Roberts & Newton, 2001; Tsuiji & Watanabe, 2010); thus, it seems reasonable to assume that was also the case in our own studies. Regardless, the data from four experiments support the conclusion that an initial judgment, however produced, produces a FOR, which, in turn, mediates the extent and outcome of subsequent Type 2 processing.

This formulation of the “extent and outcome” of Type 2 processes represents a subtle departure from early versions of DPT. Previously, it was common to talk about Type 2 processes in an “all or none” fashion, for example, that Type 2 processes would either intervene or not to change a Type 1 output (Evans, 2007a; Stanovich, 1999). Such a view is problematic in many ways because, as Kahneman (2003) argued, Type 2 processes will always be at least minimally involved in producing a response, if only because the response is made available to working memory prior to its emission. Moreover, it treats Type 2 processing as a unitary system and fails to acknowledge the range of strategies that could be engaged, only some of which will lead to an answer change and few of which will
produce a normatively correct answer (Evans, 2009; Stanovich, 2009). Thus, we would argue that multiple, converging measures should be used as an indicator of Type 2 processing.

Although the term “metacognition” is generally understood to imply conscious awareness, there are those who argue that monitoring and control processes can be initiated without such awareness (Cary & Reder, 2002; Koriat, 2007). Monitoring, on this account, is passive, with attention allocated only to a small subset of the monitored cognitions (Van Overschelde, 2008). Thus, there is no need to assume that the relationship between the FOR and Type 2 engagement is mediated by conscious awareness, any more than it is necessary to assume that reasoners are able to introspect on the origins of the FOR. Moreover, given that Type 2 processes are assumed to require working memory processes, the kind of overt monitoring that might be achieved by Type 2 processes is likely to overburden the system, leaving little capacity available for control functions (Van Overschelde, 2008).

In keeping with that idea, our experiments also provided compelling support for the role of answer fluency in monitoring Type 1 outputs. Initial answers that were generated quickly consistently produced a higher FOR than answers that were generated slowly. Thus, as in the case with memory retrievals (e.g., Benjamin et al., 1998; Jacoby et al., 1989; Kelley & Jacoby, 1993, 1996; Matvey et al., 2001; Whittlesea & Leboe, 2003), it is possible that fluency of processing engendered a positive feeling about a stimulus (see also Harmon-Jones & Allen, 2001; Winkelman, Halberstadt, Fazendeiro, & Catty, 2006; Zizak & Reber, 2004), regardless of whether participants evaluated logical validity or generated probability estimates.

These data also provide an explanation for the ubiquity of so-called reasoning biases (and the confidence with which they are held). Specifically, many of the classic reasoning problems prompt a response from automatic processes, such as linguistic comprehension, stereotyping, belief-evaluation, and imagery (Kahneman, 2003; Stanovich, 2004). The ease with which these responses come to mind may create a sense of rightness that prevents subsequent analysis (Simmons & Nelson, 2006; Thompson, 2009). That is, these processes may create an experience of fluency that is strong and that results in a high FOR, which, in turn, signals that further analysis is not required.

Although answer fluency clearly plays a role in monitoring Type 1 outputs, higher-level goals and intentional states will also play a role in initiating Type 2 thinking (Amsel et al., 2008; Klaczynski, 2006; Stanovich, 2009). For example, Stanovich (1999, 2009) has presented evidence to suggest that epistemic thinking dispositions predict the probability of generating normatively correct solutions to reasoning problems, even after accounting for variance associated with cognitive capacity. These dispositions are measured by instruments such as the Rational Experiential Inventory (Pacini & Epstein, 1999) and the Actively Open-Minded Thinking questionnaire (Stanovich & West, 2007), which assess reasoners’ self-reported tendency to prefer open-minded, analytic thinking vs. intuitive, experientially based thinking. The relationship between scores on these questionnaires and normative responses indicate that the initiation of Type 2 processes can occur at an intentional level.

The role of cognitive capacity in initiating Type 2 thinking is less clear. It is certainly the case that reasoners of higher capacity do better on a variety of reasoning tasks, at least in terms of producing normatively correct output. It is less certain, however, whether they are more successful at monitoring Type 1 outputs and thus more likely to recognize the need to engage Type 2 processes, or are simply more successful at executing Type 2 processes once they have been engaged (Evans, 2007a). Indeed, many so-called biases are insensitive to capacity (Klaczynski & Robinson, 2000; Stanovich & West, 2008; Torrens, Thompson, & Cramer, 1999), suggesting that at least for some tasks, high capacity does not guarantee good monitoring skills.

Nor, it appears, does Type 2 processing guarantee normative reasoning. In our studies, we observed that the time spent re-analysing an initial answer did not necessarily increase the probability that a normatively correct answer would be produced. Indeed, in many cases, the initial answer did not change from Time 1 to Time 2, indicating that reasoners often “went with their first instinct”. In these instances, we speculate that reasoners spend the rethinking period confabulating their initial answer (Evans, 2006; Stanovich, 2009). Alternative modes of Type 2 thinking are more cognitively demanding, requiring the reasoner to inhibit their initial answer and represent an alternative state of the
world (Evans, 2006; Handley, Capon, Beveridge, Dennis, & Evans, 2004; Simoneau & Markovits, 2003; Stanovich, 2009). In addition to this, production of a normatively correct answer requires that reasoners understand and recognise the appropriateness of a given normative rule system. In other words, relying on normative responses as an index of Type 2 thinking will almost certainly underestimate the prevalence of deliberate, analytic thought.

Finally, we found evidence for two additional determinants of the FOR. In the two experiments where participants were asked to evaluate conclusions, the FOR was higher when participants responded “yes”, i.e., that the conclusion was valid, than “no”, the conclusion was invalid. Furthermore, in both experiments, participants evidenced a decided preference to respond “yes” rather than “no”. This pattern would be promoted by a reasoning strategy focussed on consistency detection, whereby conclusions consistent with (as opposed to necessitated by) the premises are accepted (Evans et al., 1999; Thompson, Striemer, Reikoff, Gunter, & Campbell, 2003). The higher FOR associated with these “yes” responses might reflect the fact that acceptances are associated with definitive evidence, whereas rejections are associated with ambiguous evidence (i.e., failure to find sufficient grounds to accept the conclusion).

Indeed, the results of Experiment 3 also indicated that response ambiguity reduces FOR judgments. There, we observed that the FOR was highest for problems that promoted a single, unambiguous response, namely the congruent problems. In the case of the incongruent problems, the conflict between responses based on the base rates and those based on the personality descriptions may have produced ambiguity. Thus, a reasonable hypothesis arising from these findings is responses that can be made on the basis of unambiguous cues raise the FOR and ambiguity of any kind lowers it.

In summary, our data support the Metacognitive Reasoning Theory in which the quality and extent of Type 2 thinking is determined by a metacognitive experience accompanying an initial response to a problem. This metacognitive experience, in turn, appears to be multiply determined by fluency of processing and by at least one other quality, which we provisionally designate to be the ambiguity associated with the response. We have therefore provided the initial building blocks of a theory of monitoring and control processes in reasoning and decision-making.

6.2. Conclusions

DPT posits that judgments and decisions are mediated by two qualitatively different types of processes. Automatic, Type 1 processes cue an initial response to a problem that may or may not be subsequently analysed by deliberate, Type 2 processes. We have argued that current versions of these theories lack a well-articulated mechanism to determine when, and to what extent, one or the other prevails. In four experiments, we demonstrated that the monitoring and control of reasoning judgments is mediated, at least in part, by a metacognitive experience that accompanies production of an initial answer. This Feeling of Rightness is, in turn, multiply determined. These data provide an important link to theories of metacognition as well as the means to integrate two large and well-established literatures.

Acknowledgments

This research was funded by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada awarded to the first author, by a Post Graduate Scholarship from the same funding agency awarded to the second author, and by an Undergraduate Summer Research Award from the same agency to the third author. The authors would like to thank Jamie Campbell, Shira Elqayam, and Jonathan Evans for useful comments and feedback on an earlier draft of this manuscript.

References


