The Role of Answer Fluency and Perceptual Fluency as Metacognitive Cues for Initiating Analytic Thinking

Valerie Thompson, Jamie Prowse Turner

University of Saskatchewan, Canada

Gordon Pennycook,

University of Waterloo, Canada

Linden J. Ball, Hannah Brack

Lancaster University, UK

Yael Ophir, Rakefet Ackerman

Technion-Israel Institute of Technology, Israel

Running Head: Answer and Perceptual Fluency

Abstract

Although widely studied in other domains, relatively little is known about the metacognitive processes that monitor and control behaviour during reasoning and decision-making. In this paper, we examined the conditions under which two fluency cues are used to monitor initial reasoning: *answer fluency*, or the speed with which the initial, intuitive answer is produced (Thompson, Prowse Turner, & Pennycook, 2011), and *perceptual fluency*, or the ease with which problems can be read (Alter, Oppenheimer, Eply, & Eyre, 2007). The first two experiments demonstrated that answer fluency reliably predicted Feeling of Rightness (FOR) judgments to conditional inferences and base rate problems, which subsequently predicted the amount of deliberate processing as measured by thinking time and answer changes; answer fluency also predicted retrospective confidence judgments (Experiment 3b). Moreover, the effect of answer fluency on reasoning was independent from the effect of perceptual fluency, establishing that these are empirically independent constructs. In five experiments with a variety of reasoning problems similar to those of Alter et al. (2007), we found no effect of perceptual fluency on FOR, retrospective confidence or accuracy; however, we did observe that participants spent more time thinking about hard to read stimuli, although this additional time did not result in answer changes. In our final two experiments, we found that perceptual disfluency increased accuracy on the CRT (Frederick, 2005), but only amongst participants of high cognitive ability. As Alter et al.'s samples were gathered from prestigious universities, collectively, the data to this point suggest that perceptual fluency prompts additional processing in general, but this processing may results in higher accuracy only for the most cognitively able.

KEYWORDS: metacognition, reasoning, fluency monitoring and control, dual-process theories

1. Introduction

There are numerous documented situations where questions about probability or logic are answered on the basis of a readily accessible piece of information that is misleading or technically irrelevant to the decision in question (Kahneman, 2003). For example, a decision about logical validity may be made on the basis of the believability of the conclusion (Evans, Barston, & Pollard, 1983), a probability judgment may be based on a stereotype (Kahneman & Tversky, 1973), an estimate of proportion may be based on set size rather than ratio (Denes-Raj & Epstein, 1994), a decision about whether to take a risk may be based on the desirability of the outcome, rather than its probability (Finucane, Alhakami, Slovic, & Johnson, 2000). These intuitions are often accompanied by an affective experience of confirmation or sense of confidence (Hogarth, 2010; Sinclair, 2010), which may act as a disincentive to re-examine the initial answer (Thompson et al., 2011).

This paper examined potential bases for such intuitions by integrating metacognitive approaches to control of behaviour and cognitive effort with dual-process theories of reasoning. Although the various dual-process theories posit somewhat different architectures (e.g., Evans, 2007; Sloman, 1996), they nonetheless all share the basic assumption that decisions are based on the output of the relatively faster, more automatic Type 1 processes, unless they are overturned or changed by more deliberate, analytic Type 2 processes. For example, on the default interventionist view (Evans, 2006; Kahneman, 2003; Sloman, 1996; Stanovich, 2004), Type1 processes deliver an initial answer, which may or may not be scrutinized by analytic, Type 2 processes. In contrast, on the parallel competitive view, both Type 1 and Type 2 are engaged from the outset, but the output from the faster Type 1 processes often form the final answer

(Sloman, 1996). A fundamental goal for these theories, regardless of their stance on the time course of analytic engagement, is to understand conditions under which additional Type 2 processing occurs. For this reason, several researchers have argued that dual-process theories need to incorporate a third category of processes that monitors Type 1 outputs and that initiates Type 2 analysis (De Neys & Glumicic, 2008; Evans, 2009; Simmons & Nelson, 2006; Stanovich, 2009; Thompson, 2009, 2010) .

Thompson (2009, 2010) suggested that theorizing about the link between Type 1 and Type 2 processes should begin with the already well-established literature on metacognition, where the distinction between the processes responsible for retrieving information from memory and the processes responsible for monitoring that information is well documented. Although it is possible for monitoring to be based on explicit knowledge, such as beliefs about one's skill at a task (e.g., Dunning, Johnson, Ehrlinger, & Kruger, 2003; Prowse Turner & Thompson, 2009) or one's lay theories about cognitive functions (Koriat, Bjork, Sheffer, & Bar, 2004), most of the theorising has focussed on the role of implicit cues, such as the ease with which an item is retrieved from memory (Benjamin, Bjork, & Schwartz, 1998; Koriat & Ma'ayan, 2005; see Koriat, 2007 for a review). In turn, the resultant metacognitive monitoring plays a causal role in determining subsequent behaviour (e.g., Ackerman & Goldsmith, 2008; Metcalfe & Finn, 2008; Singer & Tiede, 2008; Son, 2004), even among young children (Koriat & Ackerman, 2010).

1.1. Answer fluency, metacognition, and Type 2 thinking

In terms of reasoning, Thompson (2009) posited that the processes that monitor Type 1 outputs might be sensitive to the ease with which that response was generated, a variable termed *answer fluency* (see also Simmons & Nelson, 2006). Specifically, she proposed that the fluency with which Type 1 processes produce an initial answer gives rise to a metacognitive judgment, called the Feeling of Rightness (FOR) that in turn mediates the extent of Type 2 engagement. That is, fluently generated outputs are postulated to create a strong FOR, which signals that elaborated processing is not required (in default interventionist models, e.g., Evans, 2006; Kahneman, 2003) or that Type 2 processing may now cease (in parallel processing models, e.g., Sloman, 1996). On this account, the reason that many reasoning "biases" are so compelling and persistent may be because they are highly accessible (Kahneman, 2003; Stanovich, 2004) and this ease of retrieval creates a strong sense of rightness (Simmons & Nelson, 2006; Thompson, 2009; Topolinski & Reber, 2010).

To test this hypothesis, Thompson et al. (2011) asked participants to give a fast, intuitive answer to a reasoning problem, rate their FOR, and then take as much time as needed to produce a final answer. Over a series of problems, it was observed that answers that were generated fluently (quickly) were accompanied by a strong FOR, whereas less fluently retrieved answers were associated with weaker FORs. In turn, the strength of the FOR predicted the extent of Type 2 engagement, as indexed by thinking time and the probability of changing an answer.

1.2. Perceptual fluency, metacognition, and Type 2 thinking

Working from a different tradition, Alter, Oppenheimer, Eply, and Eyre (2007) proposed that making problems difficult to perceive can trigger Type 2 thinking, a variable we call *perceptual fluency*. They found that reasoners were more likely to correctly solve Frederick's (2005) Cognitive Reflection Test (see the Appendix) and three-term syllogisms when they were presented in a difficult to read font than when the font was easy to read; similarly, when people were cued to interpret their efforts as disfluent by furrowing their brow, they were more likely to give normatively correct answers to base rate problems. Alter and Oppenheimer (2009) argued that the disfluency produced by the difficult-to-read font generated a metacognitive cue, which signalled that further analysis was required. When information was processed fluently, Type 1 processing was hypothesized to dominate. This hypothesis is consistent with the abundant evidence to suggest that fluently processed items are perceived more positively than those requiring more effort (see Alter & Oppenheimer, 2009; Topolinksi & Reber, 2010 for recent reviews), such that, for example, fluently processed statements are judged true more often than their disfluent counterparts (Reber & Schwarz, 1999).

Although answer fluency and perceptual fluency are often treated as variables that load onto the same metacognitive construct, (Alter & Oppenheimer, 2009; Briñol, Petty & Tormala, 2006), in the current work, we emphasize their distinctiveness (Hertwig, Herzog, Schooler & Reimer, 2008). Specifically, perceptual fluency affects the ease with which an entire problem or question is experienced, whereas answer fluency specifically refers to the ease with which a particular answer comes to mind. Thus, for a given level of perceptual 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. Conversely, both answers that are produced fluently and disfluently may be subject to the same feeling of unease if the problems that produced them are difficult to process.

2. Experiments 1a and 1b: Perceptual and Answer Fluency as Metacognitive Cues

The goal of these first two studies was to test the relationship between perceptual fluency, answer fluency, FOR, and Type 2 thinking. A second goal was to establish empirically that answer fluency and perceptual fluency are independent cues to metacognitive judgment. To do so, we extended Thompson et al.'s Feeling of Rightness (FOR) analysis to Alter et al.'s (2007) perceptual fluency manipulation, under the hypothesis that the metacognitive unease produced by the perceptual disfluency would be reflected in lower FOR judgments. Thus, in both experiments, participants provided two answers to each problem: An intuitive fast response and a final response given without time constraint. FORs were measured after providing the intuitive fast response. In Experiment 1a, perceptual fluency was manipulated by presenting the problems to half of the participants in a difficult to read font and to the other half in an easy to read font, following Alter et al.'s Experiments 1 and 4. In Experiment 1b, perceived fluency was manipulated by asking half of the participants to solve the problems while furrowing their brows and half while puffing their cheeks, following Alter et al.'s Experiment 3. The difficult to read font and the furrowed brows were found by Alter et al. to increase the probability of normatively correct responses.

Answer fluency was measured by the time it took to provide the intuitive fast response. To measure response time accurately, the problems were presented on a computer.[1] In order to make the results of the two sets of studies comparable, we used a wide range of measures of Type 2 processing. Specifically, whereas Thompson et al. (2011) measured time spent rethinking after providing the initial answer and the probability that this answer would be changed, Alter et al. used the traditional measure of normative correctness as a benchmark of Type 2 processing. To make the studies comparable, we used all three measures in these studies.

[1] We note that Alter et al. presented problems using pencil and paper. To rule out media differences as factor in our data, a pilot study was conducted with 114 undergraduate students (*N* = 114). They solved the Cognitive Reflection Test (Alter et al., 2007, Experiment 1) presented in difficult or easy to read fonts, either on a computer screen or on paper. No media effect and no interaction with the font type were found, both *F*s < 1.

2.1. Experiment 1a

Following Thompson et al. (2011, Experiment 1), participants judged the validity of conclusions drawn from familiar conditional relations. The problems were presented either in difficult or easy to read font. At first we chose fonts to be as similar as possible to those described in Alter et al. (2007); when we observed the fonts were not having the anticipated effect, we increased the challenge posed by the difficult-to-read font. Under the hypothesis that the two types of fluency would be independent, we predicted that the FORs would be lower for the difficult than for the easy to read font, and that within each font type, quickly generated answers would produce higher FORs than answers that took longer to produce. Moreover, we predicted that higher FORs would produce less Type 2 thinking.

2.1.1. Method

2.1.1.1. Participants

One hundred and eight introductory psychology students at the University of Saskatchewan (76% female, mean age = 19.5 years) received partial course credit for their participation.

2.1.1.2. Materials

Sixteen conditional statements (from Thompson et al., 2011) were presented with a minor premise and conclusion conforming to four inferences, yielding 64 problems. The minor premises asserted the occurrence or non-occurrence of the antecedent or consequent events. The participants were asked to draw a conclusion about the other event, as illustrated below:

If the car is out of gas, then it stalls;

The car is out of gas. Therefore, it stalls.

Yes/No

Half of the conclusions were believable (as above) and half were unbelievable (e.g., If a plant has roots, then it is a tree. This plant has roots. Therefore, it is a tree), as established by prior ratings (Thompson, 1994). On separate trials, the two valid and two invalid inferences were presented for each conditional; validity varied orthogonally to believability. Given that the effects of these variables on reasoning performance are well-studied, they will not be reported here.

2.1.1.3. Procedure

Prior to testing, two practice problems were presented. The problems were presented one at a time on a computer screen followed by two response options. Participants were instructed to choose "yes" if the conclusion followed logically from the premises and "no" if it did not. Problems were presented in a different random order for each participant. The response keys marked "yes" and "no" alternated across participants.

In the perceptually fluent condition, problems were presented in 10 point Courier New black font on a white background. For 30 participants in the disfluent condition, problems were presented in italicized light grey font, in 7 point Courier New font (similar to Alter et al.). When it appeared that indices of Type 2 thinking did not differ between the two groups, we increased the difficulty of reading the font in the disfluent condition by presenting the text in teal italicized 10 point Curlz MT font on a green background[2].

All participants were instructed to give two responses to each problem. The first was to

[2] This combination was chosen on the basis of a pilot study that showed it to be particularly difficult to read but still legible.

be the first response that came to mind; intuitive responding was emphasized. Timing started with the presentation of the problem and stopped when they made their first response. They gave their FOR on a seven-point Likert scale ranging from "Guessing" (1) to "Certain I'm right" (7). 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?[3]" The problem then re-appeared and participants were allowed as much time as they wanted to give their second answer[4]. Timing started when the problem re-appeared. Participants were instructed to take their time and make sure that they gave the correct answer.

2.1.2. Results

For each problem, we measured the time to generate the first response (answer fluency), accuracy for the first response, Feeling of Rightness (FOR) rating, time to generate the second response (rethinking time), probability of changing answers between first and second response, and accuracy for the second response. RTs for this and all subsequent experiments were converted to $\log^{10}$ prior to analysis (RTs in the Tables are reported in the original units). The data are summarized in Table 1.

2.1.2.2. Perceptual fluency, FOR, and Type 2 processing

Performance in the perceptually fluent and disfluent conditions were virtually identical[5].

[3] Note that five participants were replaced because they answered "no" to more than 10% of the trials. On average, participants indicated that they responded with the first answer that came to mind on more than 98% of trials.

[4] Participants also provided a final judgment of confidence, which did not differ between the two fluency conditions, $t < 1$.

[5] When we included our two difficult-to-read fonts as a variable in the analysis, none of the interactions or main effects involving this variable were significant (largest $F = 2.30$, $p >$

Table 1. *Reasoning performance as a function of perceptual fluency manipulation in Experiment 1a; standard errors are in parentheses.*

		Perceptual Fluency	
Dependent Measure		Easy	Difficult
First Response			
	RT (sec)	7.2(.29)	7.5(.34)
	Accuracy	0.53(.01)	0.54(.01)
	FOR (max = 7)	5.43(.11)	5.49(.12)
Second Response			
	RT (sec)	6.4(.52)	5.6(.52)
	Probability Change	0.12(.01)	0.10(.01)
	Accuracy	0.55(.01)	0.55(.01)
FOR			
	Faster First Response	6.13(.10)	6.20(.10)
	Slower First Response	5.84(.11)	5.84(.11)
	Faster Second Response	5.68(.12)	5.78(.12)
	Slower Second Response	5.18(.13)	5.20(.13)
	Answer Changed	4.44(.15)	4.21(.15)
	Answer Not Changed	5.55(.12)	5.57(.12)
	Answer Correct	5.42(.12)	5.50(.12)
	Answer Incorrect	5.45(.12)	5.24(.12)

.13), thus, for the sake of simplicity, they were combined.

The difficult to read font did not engender lower FORs than the easy to read one, nor did any of the indicators of Type 2 thinking (i.e., rethinking time, probability of changing answers, or probability of correct answers) differ between the two font types, $t(106) \leq 1.20, p \geq .23$.

2.1.2.2. Answer fluency, FOR, and Type 2 processing

To test the relationship between answer fluency and FOR, we computed, for each participant, the median time to generate the first response. Answers falling above the median were designated as "slower," representing disfluent answering, and those falling below the median were classified as "faster," representing fluent answering. We then computed the mean FOR for fluent and disfluent responses; similarly, we computed the mean FOR for long and short rethinking times and for answers that were changed or remained the same. The results were analysed using 2 (perceptual fluency: font type) x 2 (faster/slower answers) mixed ANOVAs (see Table 1).

As expected, FORs were higher when the first answer was produced quickly than when it was produced slowly, $F(1,106) = 119.4, p < .001, \eta_p^2 = .53$; neither the main effect of perceptual fluency nor the interaction were reliable, $F(1,106) \leq 1.51, p \geq .22, \eta_p^2 \leq .01$. Also as expected, FORs were associated with two measures of Type 2 thinking, such that weak FORs were followed by more Type 2 thinking than strong ones: The FORs preceding long rethinking times were lower than the FORs preceding short ones, $F(1,106) = 119.1, p < .001, \eta_p^2 = .53$ and answers that changed were preceded by lower FORs than answers that were not changed, $F(1,102) = 167.3, p < .001, \eta_p^2 = .62$. In neither case was the main effect of perceptual fluency or the interaction significant, $F(1,102) \leq 1.70, p \geq .195, \eta_p^2 \leq .02$. A similar relationship was not observed between FORs and accuracy, $F < 1$.

2.1.3. Conclusions

The findings regarding answer fluency are consistent with the hypothesis that intuitive responses are accompanied by a metacognitive judgment (FOR) that determines the extent of Type 2 thinking. The strength of this feeling of rightness varies as a function of the ease with which the initial answer comes to mind: Fluently generated answers engender a stronger FOR than their less fluent counterparts. In turn, this FOR signals the need (or lack thereof) for additional Type 2 processing: relative to weak FORs, strong FORs are associated with shorter rethinking times and lower probability of answer changes. Note, however, that our third measure of Type 2 processing, i.e., probability of correct responding, was not sensitive to FOR judgments.

In that regard, our findings are consistent with those who argue that normative accuracy is not necessarily a good index of Type 2 thinking (Elqayam & Evans, 2011), given that a) normatively correct responses can be achieved using non-analytic heuristics (Gigerenzer, Todd, & the ABC Reasoning Group, 1999; Oaksford & Chater, 2007), and b) that Type 2 thinking may not produce a normatively correct response (Evans, 2006; Stanovich, 2009). That is, even when an alternative is sought, knowledge of the relevant normative principles (e.g. of logic or probability) is required to produce a correct answer. Thus, it is easy to observe markers of Type 2 thinking, such as additional rethinking times and answer changes, in the absence of an effect on normative accuracy.

As predicted, the relationship between FOR and Type 2 processing did not interact with processing fluency, consistent with our hypothesis that these are independent constructs. That is, regardless of whether the stimuli were easy or difficult to read, answers that were retrieved fluently engendered higher FORs and less Type 2 thinking than their less fluent counterparts.

We note also that the manipulation of perceptual fluency did not affect answer fluency, such that RTs for the initial responses were similar in the two conditions, which is again consistent with the hypothesis that these are independent constructs.

We had, however, anticipated that both perceptual and answer fluency would provide independent bases for FOR and act as two cues to Type 2 thinking. Surprisingly, we did not observe the expected relationship between perceptual fluency and either FOR judgments or Type 2 engagement. That is, unlike Alter et al. (2007), we were unable to demonstrate that perceptual disfluency allowed reasoners to overcome their initial intuition in favour of a correct answer.

2.2. Experiment 1b

In their Experiment 3, Alter et al. (2007) manipulated a variable designed to affect the feeling of effort (see Koriat & Nussinson, 2009; Stepper & Strack, 1993). They did so by asking half their participants to furrow their brows while completing a single probability estimation task (Kahneman and Tversky's, 1973 "Tom W" problem), whereas the other half puffed their cheeks. Consistent with their hypothesis, Alter et al. (2007) observed that participants' probability estimates were more likely to integrate base rate information with the diagnostic information in the high effort (furrowed brow) than the low effort (puffed cheek) facial expression.

Here we employed the variant of the Tom W. problem where participants are given a number of problems where base-rate information is paired with stereotypes (De Neys & Glumicic, 2008; Thompson et al., 2011). Following Alter et al., half the participants furrowed their brows while solving the problems and half puffed their cheeks. As before, we predicted that FORs should be lower and signatures of Type 2 thinking should be higher in the high effort than the low effort facial expression. We also predicted that answers that were produced fluently would engender higher FORs and less analytic thinking than their less fluent counterparts. Finally, these variables were predicted to be independent, such that the effect of answer fluency should be observed regardless of facial expression .

2.2.1. Method

2.2.1.1. Participants

Seventy-two University of Saskatchewan undergraduate students (78% female, mean age = 20.6 years) received partial course credit for an introductory psychology class (*N* = 48) or were recruited from advertisements and were paid CAN $5.00 (*N* = 24).

2.2.1.2. Materials

These were identical to Thompson et al.'s (2011) Experiment 3. There were 18 probability estimation problems adapted from De Neys and Glumicic (2008) plus one practice problem. Each problem contained information about the base rate of belonging to a particular group along with a personality description of an individual. The base rate and personality descriptions either provided congruent, incongruent, or neutral information (there were six of each type). Examples of each type appear below:

Incongruent:

> In a study 1000 people were tested. Among the participants there were 997 nurses and 3 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?

Congruent:

In a study 1000 people were tested. Among the participants there were 996 kindergarten teachers and 5 executive managers. Lilly is a randomly chosen participant of this study. Lilly is 37 years old. She is married and has 3 kids. Her husband is a veterinarian. She is committed to her family and always watches cartoons with her kids.

What is the probability that Lilly is a kindergarten teacher?

Neutral:

In a study 1000 people were tested. Among the participants there were 995 who live in Los Angeles and 5 who live in New York. Christopher is a randomly chosen participant of this study. Christopher is 28 years old. He has a girlfriend and shares an apartment with a friend. He likes watching basketball.

What is the probability that Christopher lives in Los Angeles?

Two versions of the stimulus set were created wherein each stereotypical personality description matched the larger group (congruent) or smaller group (incongruent) an equal number of times. Three extreme base-rate probability ratios were presented equally often in each congruency condition; 995/5, 996/4, 997/3. Extreme base-rates were used in order to maximize conflict detection and, by extension, Type 2 engagement (Pennycook, Fugelsang, & Koehler, 2012). For each description, the large category was asked about half the time, and the small category was asked about half of the time. The order of problems was randomly determined for each individual.

2.2.1.3. Procedure

The problems were presented one at a time on a computer screen. Participants typed their answers. The two-response procedure was identical to that used in Experiment 1, except that participants were told to make their first response under a deadline (12 sec), as per Thompson et al. (2011)[6]. The study was titled "Nonverbal Emotional Expressions and Reasoning", and participants were told that they would be asked to hold a certain facial expression while doing the reasoning task. The facial expression (furrowed brow or puffed cheeks) was then demonstrated to the participant, and the participant had to adopt the expression (showing the researcher that they are able to do so). The experimenter monitored the participants throughout the experiment to ensure that they complied with the instructions, although he was not in their line of vision.

2.2.2. Results

2.2.2.1. Scoring

The scores for items that asked about the smaller of the two groups (e.g., there were 995 doctors and 5 nurses; what is the probability that Paul is a nurse?) were subtracted from 100. Thus, high scores always indicated estimates that were close to the base-rate and low numbers always reflected estimates that deviated from the base-rate. We discarded responses for which participants indicated that they did not respond with the first answer that came to mind. The dependent measures were: time to generate the first response (answer fluency), probability estimates at Time 1[7], FOR rating, amount of time to generate the second response (rethinking time), degree of change in probability estimates between first and second response, probability

[6] One participant was replaced for indicating that he/she did not respond with the first answer on more than one of the 18 trials. The remaining participants indicated that they responded with their first answer on 96% of trials.

[7] Note that it is not possible to compute accuracy on this task without an estimate of the diagnosticity of the personality description.

estimates at Time 2[8]. All RTs were converted to $\log^{10}$ prior to analysis (RTs in the Table are reported in the original units).

2.2.2.2. Facial expression, FOR, and Type 2 processing

We adopted the same analysis strategy as in Experiment 1a and began by analysing the effect of the facial expression manipulation; these data are summarized in Table 2. The data were analysed using 2 (facial expression) x 3 (congruency) mixed ANOVAs. However, because the effects of congruency were secondary to our interest, and because they replicated our previous work (Thompson et al. 2011), they will not be reported here; congruency did not interact with facial expression for any of the analyses, all *Fs* < 1.

Once again, performance in the two facial expression conditions was virtually identical, with one notable exception. Specifically, the furrowed brows did not engender a lower FOR than the puffed cheeks, $F(1, 70) = 2.64$, $MSE = 2.71$, $p > .1$, $\eta_p^2 = .04$, first responses were not generated more quickly in the puffed cheeks condition, $F(1, 70) = 1.24$, $MSE = .04$, $p = .27$, $\eta_p^2 = .02$, the degree of answer change was similar in the two facial expressions, as were probability estimates at both Time 1 and Time 2, all *F*s < 1. However, there was some evidence of additional Type 2 thinking in the furrowed brow condition, in that reasoners took longer to rethink their answers than in the puffed cheeks condition, $F(1, 70) = 8.27$, $MSE = .21$, $p = .005$, $\eta_p^2 = .11$.

2.2.2.3 Answer fluency, FOR, and Type 2 thinking.

The relationship between FOR, answer fluency, rethinking time, and answer change was identical to that reported in Experiment 1a. Again, we computed, for each participant the median

[8] Again, participants provided final judgments of confidence after their final answers, which, as in Experiment 1, did not differ between facial expressions, $t < 1$.

Table 2. *Reasoning performance as a function of facial expression in Experiment 1b; standard errors are in parentheses.*

	Facial Expression	
Dependent Measure	Puffed Cheeks	Furrowed Brows
First Response		
RT (sec)	14.6(.66)	15.6(.66)
Probability Estimate	68.7(2.63)	67.6(2.63)
FOR (max = 7)	4.96(.16)	4.58(.16)
Second Response		
RT (sec)	12.6(1.18)	17.1(1.18)
Probability Estimate	71.4(2.90)	73.8(2.90)
Degree of Change	13.87(1.57)	14.98(1.57)
FOR		
Faster First Response	5.10(.16)	4.70(.16)
Slower First Response	4.81(.16)	4.46(.16)
Faster Second Response	5.22(.18)	4.77(.18)
Slower Second Response	4.69(.17)	4.40(.17)
Answer Changed	4.49(.16)	4.27(.16)
Answer Not Changed	5.20(.18)	5.08(.17)

time to generate the first response and examined whether there are FOR differences in light of this division. These data are also presented in Table 2. The data were analysed using 2 (facial expression) x 2 (answer fluency) mixed ANOVAs. As expected, FORs were higher when the initial answer was generated quickly than when it was provided after longer times, for short rethinking times than long ones, and for answers that did not change relative to those that did, $F(1,70) \geq 26.9, p < .001, \eta_p^2 \geq .28$. None of the main effects of facial expression nor the interaction were reliable, $F(1,70) \leq 2.75, p \geq .10$.

2.2.3. Conclusions

As was the case for Experiment 1a, the data from Experiment 1b are consistent with the hypothesis that initial responses have two dimensions (Thompson et al., 2011): The first is the answer itself and the second is a metacognitive judgment that accompanies that answer. The strength of this feeling of rightness varies as a function of the speed or fluency with which the initial answer comes to mind, and the strength of this FOR predicts Type 2 engagement. As was the case with perceptual fluency, the facial expression manipulation did not affect the fluency of the initial answer, FORs, or answer changes. In contrast, we did observe longer rethinking times in the furrowed brow condition, suggesting that participants in the furrowed brow condition were engaging in more Type 2 thinking than their counterparts in the puffed cheek condition. However, this additional thinking had little effect on their responses, given that both the degree of change and their final probability estimates were similar. We posit that the furrowed brow condition prompted a sufficient sense of unease to promote additional rethinking time, however, this additional processing was not necessarily “deeper” processing, in that it did not prompt reasoners to change their answers. It should be noted that this can not be explained by an overall lack of answer change, as over 70% of answers were changed.

3. Experiments 2a, 2b, and 2c: Perceptual fluency and deductive reasoning

Our inability to reproduce Alter et al.'s (2007) perceptual fluency effects on normative responding is puzzling, although we note that there were several methodological differences. For example, our participants gave two answers and FOR judgments, whereas their participants provided a single answer[9]. In addition, although our stimuli were conceptually similar to Alter et al.'s there were also potentially important differences. First, instead of syllogisms, we used conditional statements that produced mean performance marginally above chance, suggesting a limited normative basis to participants' responses. Second, whereas Alter et al. used a single base-rate problem, we presented multiple base-rate problems using similar probability information, which may have cued Type 2 thinking (Kahneman, 2000), thereby obscuring a perceptual fluency effect. In an effort to replicate Alter et al.'s (2007) findings, we conducted three experiments that adopted similar tasks and methods to theirs. Table 3 provides a summary of key methodological aspects of all three experiments.

3.1. Experiment 2a

Experiment 2a used a between-participants fluency manipulation, with participants solving belief-neutral syllogisms, as per Alter et al. (2007, Experiment 4). Predictions are summarised in Table 3 and performance data (mean accuracy and RTs) are presented in Table 4. Data were analysed using 2 (perceptual fluency) x 2 (validity) mixed ANOVAs.

Although reasoners were more accurate with valid than invalid problems

[9] However, see Thompson et al. (2011) for evidence that providing the first answer and FOR judgment does not change the answer given at Time 2.

Table 3. M*ethodological aspects of Experiments 2a, 2b and 2c, which aimed to replicate the disfluency effect observed by Alter et al. (2007).*

Experiment	**Task type**	**Perceptual fluency manipulation**	**Predictions**	**Materials**	**Procedure**
2a (N=48, University of Lancaster undergraduates)	Belief-neutral syllogisms with valid and invalid conclusions.	A similar font manipulation was used to Alter et al. Participants received syllogisms in either an easy-to-read font (Arial 18-point black) or a hard-to-read font (Brush script 18-point italicised light grey). Pre-testing established that font legibility was significantly polarised, although the hard-to-read font was still legible.	Disfluent presentation should cue Type 2 processing, i.e., improved valid vs. invalid discrimination and longer response latencies relative to fluent presentation.	Valid and invalid syllogisms (8 of each) such as: Some butchers are skaters; No skaters are painters; Therefore, some butchers are not painters.	Computer-presented problems in a randomised order, with a single accept/reject decision for each given conclusion. Processing time measures were taken for all premises and conclusions using Stupple and Ball's (2008) inspection-time methodology.
2b (N=42, University of Lancaster undergraduates)	Belief-laden syllogisms with valid and invalid conclusions that were either believable or unbelievable.	An identical font manipulation to Experiment 2a.	Disfluent presentation should promote greater reliance on validity (arising from Type 2 processing) and less reliance on belief (arising from heuristic processing) relative to fluent presentation.	Equivalent materials to Experiment 2a, except that half of the syllogisms had believable conclusions (e.g., Some fruit are not apples) and half had unbelievable conclusions (e.g., Some sparrows are not birds).	A near-identical procedure to Experiment 2a, with participants instructed to assume the truth of all premises irrespective of believability.
2c (N=287; 48 completed the CRT after Exp 1a; 239 were recruited from local Canadian websites (Kijiji)).	Cognitive Reflection Test (CRT; Frederick, 2005).	For 48 participants the font manipulation was as described in Experiment 1. For the remainder (tested on-line) it was as employed by Alter et al. (2007, Experiment 2), with disfluency arising from the use of alphanumeric characters (e.g., β@t & β@ll = bat and ball).	Disfluent presentation should promote greater reliance on Type 2 processing, leading to enhanced solution success.	The CRT involves three simple problems that elicit *intuitive*, but incorrect, initial responses (see The Appendix).	The three problems were always presented in the same order (bat & ball, widget, lily pad). Note that 48 participants completed the CRT after completing Experiment 1a.

Table 4. *Reasoning performance (mean accuracy and mean response times) as a function of fluency condition for Experiment 2a; standard errors are in parentheses.*

	Fluent	Disfluent	Mean
Validity			
	Accuracy		
Valid	.82 (.04)	.84 (.04)	.83 (.03)
Invalid	.39 (.05)	.43 (.07)	.41 (.04)
Mean	.61 (.04)	.64 (.05)	
	Response Time		
Valid	19.65 (1.75)	26.77 (1.86)	23.21 (1.37)
Invalid	21.51 (1.76)	31.35 (3.19)	26.43 (1.94)
Mean	20.58 (1.24)	29.06 (1.86)	

, $F(1,46) = 75.55$, $MSE = 0.06$, $p < .001$, $\eta_p^2 = .62$, accuracy did not differ between fluency conditions, $F < 1$. The interaction between these variables was also not significant. We conducted an additional analysis to provide closer parity to Alter et al. (2007), who discarded the easiest and hardest items to eliminate ceiling and floor effects, reporting accuracy only for moderately difficult items. The results of this analysis were equivalent to those seen with the full data-set.

RTs were converted to $\log^{10}$ prior to analysis (RTs in Table 4 are in original units). Invalid problems had longer RTs than valid problems, $F(1,46) = 5.25$, $MSE = .01$, $p = .027$, $\eta_p^2 = .10$ and problems in the disfluent condition had longer RTs than those in the fluent condition, $F(1,46) = 10.86$, $MSE = .06$, $p = .002$, $\eta_p^2 = .19$. The problem inspection-time methodology that we implemented allowed us to extract a measure of 'thinking time' by removing the time that participants spent on the initial reading of premises and conclusions from the analysis. The fluency effect identified on overall RTs remained significant for this thinking-time measure, $F(1,46) = 5.41$, $MSE = .11$, p = .024, $\eta_p^2 = .10$.

The accuracy data for Experiment 2a are again inconsistent with Alter et al.'s findings. Whereas they observed higher rates of correct responding in the disfluent than the fluent condition, we observed that accuracy in these two conditions was virtually identical. RTs, however, did support the hypothesis that disfluent processing engenders Type 2 thinking, in that reasoners spent longer thinking about problems in the disfluent font. As in Experiment 1b and the previous Experiment, longer thinking times did not increase solution accuracy.

3.2 Experiment 2b

So far we have reported two experiments measuring deductive reasoning that were unable to demonstrate beneficial effects of perceptual disfluency on accuracy. We therefore conducted a final deductive reasoning experiment designed to maximise the potential for perceptual disfluency to affect the outcome of Type 2 processing. We did this by adding conclusion believability as a variable, which is known to motivate Type 2 thinking (Evans, Handley, & Harper, 2001; Klauer, Musch, & Naumer, 2000; Stupple & Ball, 2008; Thompson, Striemer, Reikoff, Gunter, & Campbell, 2003). Predictions are summarised in Table 3.

Performance data (mean accuracy and RTs) are presented in Table 5. Data were analysed using 2 (validity) x 2 (believability) x 2 (perceptual fluency) mixed ANOVAs. The accuracy data replicated standard effects of belief and validity (Evans et al., 1983): reasoners were more accurate with valid than invalid conclusions and with unbelievable than believable conclusions, $F(1,40) = 20.47$, $MSE = 0.11$, $p < .001$, $\eta_p^2 = .34$, and $F(1,40) = 17.12$, $MSE = 0.07$, $p < .001$, $\eta_p^2 = .30$, respectively. The standard interaction between validity and believability also emerged, $F(1,40) = 13.14$, $MSE = 0.08$, $p < .001$, $\eta_p^2 = .25$, with especially poor performance arising for the invalid-believable problems. In terms of perceptual fluency, performance was comparable for the fluent and disfluent conditions, $F(1,40) = 1.56$, $MSE = 0.09$, $p = .22$, $\eta_p^2 = .04$, which remained the case even when we discarded the easiest and hardest items from each belief x validity cell. There was also no three-way interaction.

RTs were converted to $\log^{10}$ prior to analysis (Table 5 presents RT data in original units). Invalid problems had longer RTs than valid problems, $F(1,40) = 14.55$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = .27$, and a similar separation arose for believable versus unbelievable problems, $F(1,40) = 11.78$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = .23$. RTs were also significantly longer in the disfluent than the fluent condition, $F(1,40) = 4.07$, $MSE = 0.01$, $p = .05$, $\eta_p^2 = .09$. No other effects were

Table 5. *Reasoning performance (mean accuracy and mean response times) as a function of fluency condition for Experiment 2b; standard errors are in parentheses.*

Believability	Validity	Fluent	Disfluent	Mean
		Accuracy		
Believable				
	Valid	.85 (.04)	.69 (.06)	.77 (.04)
	Invalid	.48 (.09)	.36 (.08)	.38 (.05)
	Mean	.62 (.05)	.53 (.05)	
Unbelievable				
	Valid	.76 (.08)	.80 (.07)	.78 (.05)
	Invalid	.75 (.05)	.67 (.07)	.71 (.04)
	Mean	.76 (.05)	.73 (.05)	
		Response Time		
Believable				
	Valid	19.29 (2.06)	22.06 (2.06)	20.67 (1.46)
	Invalid	23.15 (2.06)	27.86 (3.29)	25.51 (1.95)
	Mean	21.22 (1.47)	24.96 (1.97)	
Unbelievable				
	Valid	15.70 (1.50)	21.49 (1.78)	18.59 (1.24)
	Invalid	18.90 (2.00)	23.36 (1.90)	21.03 (1.40)
	Mean	17.30 (1.26)	22.38 (1.29)	

significant. As in Experiment 2a, the time-based fluency effect was also significant when we analysed thinking times after first removing initial reading times for premises and conclusions, $F(1,40) = 6.25$, $MSE = 0.44$, $p = .017$, $\eta_{p2} = .14$.

Again, the accuracy data for Experiment 2 failed to reveal any benefits arising from perceptual disfluency. However, as was the case with the previous two experiments, there was evidence to suggest that disfluency promoted a reliable increase in RTs and thinking time relative to the fluent condition. To the extent that thinking time indexes the effort devoted to a problem, it appears that perceptual disfluency promotes Type 2 thinking; however, it is clear that additional thinking time does not necessarily produce differences in responses.

3.3 Experiment 2c

One explanation for the failure to find such differences may be that our sample sizes were insufficient to observe a disfluency effect. For example, with 41 participants, Alter et al. (2007) reported $\eta_p^2 = .09$ for the disfluency effect on syllogisms; Experiments 2a and 2b using 48 and 42 participants respectively had power = .55 and .49 to detect a similar effect size[10]. We note, however, that our Experiments 1a and 1b had adequate (.8) power to detect effect sizes equal to .07 and .10, which were comparable in size to those reported by Alter et al. Nonetheless, if the perceptual fluency manipulation had a smaller effect size for the particular tasks we used, then we would require correspondingly larger sample sizes to observe the effect. To address this concern we used an identical task to Alter et al. (Experiment 1), that is, the CRT and we also tested a very large group of participants (see Table 3 for methodological details). For the

[10] Power calculations performed using MorePower 6: https://wiki.usask.ca/pages/viewpageattachments.action?pageId=420413544

participants who solved the CRT after Experiment 1a, performance was the same in the fluent (*M* = .78/3) and disfluent conditions (*M* = .96/3), $t < 1$, as was that of the online participants, *M* = 1.26/3 and 1.13/3, respectively, $t < 1$. Sample size cannot be an explanation for the findings, given that the effect was not observed even when using a very large sample.

4. Experiments 3a and 3b Perceptual fluency and cognitive ability

In our first two studies, we found that perceptual fluency did not affect metacognitive judgments or the probability that reasoners changed their answers. In five experiments, we have failed to observe any evidence that creating disfluent processing conditions produces more correct solutions to similar problems. However, in three of four experiments where solution times were measured (Experiments 1b, 2a, and 2b, but not Experiment 1a) we noted that participants thought about the problems longer in the disfluent conditions, suggesting that the disfluent condition triggered some form of Type 2 thinking, albeit one that did not induce participants to change their answers.

As we argued earlier, using normatively correct answers may limit one's ability to detect Type 2 thinking. Stanovich (2009) and Evans (2007) have suggested that producing a normatively correct solution requires the reasoner not only to engage in analytic thinking, but also to have the ability to inhibit their first answer and to formulate an alternative; in addition, they must have access to the relevant rules of logic or probability and recognize the appropriateness of applying them. Thus, better educated reasoners of high cognitive capacity are more likely to produce normatively correct outcomes when they engage in Type 2 thinking (Evans, 2007; Stanovich, 1999).

On this view, the reason that Alter et al.'s manipulations produced the expected outcomes is that they conducted their experiments on an elite group of participants (Princeton and Harvard undergraduates) who had the requisite skills to produce normative answers (Brase, Fiddick, & Harries, 2006). Indeed, Frederick (2005) found a correspondence between the prestige of the institute and success rates in the CRT problems. Note that Alter et al. reported performance on the CRT ranging from 1.90/ 3 to 2.5/3, substantially higher than performance in our Experiment 2c (about 1/3). A mean of 1.90 is higher than all but one of the 11 samples gathered by Frederick (2005); only students from the Massachusetts Institute of Technology scored better at 2.18. Mean performance across Frederick's samples was 1.24. Frederick's mean performance for a web-based sample (1.10) was similar to ours; our student sample were similar to those drawn from the University of Michigan (M = .83) and Michigan State University (M = .79). Thus, it seems probable that the sample used by Alter et al. was particularly gifted relative to ours.

The following two experiments examine the hypothesis that the perceptual fluency manipulation produces correct answers only for the most able reasoners because they have the requisite skills for their Type 2 thinking to produce normative outcomes. Some support for this hypothesis comes from Cokely and colleagues (Cokely, Parpart, & Schooler, 2009) whose fluency manipulation had a greater impact on those with high CRT scores. Although suggestive that capacity mediates the fluency effect, they did not include a direct measure of cognitive capacity and the problems they used were judgments under uncertainty, compared to the complex reasoning tasks used by us and by Alter et al. Finally, our interest is in the converse effect of fluency on CRT performance; in the next two experiments, we test the hypothesis that the effect of fluency on accuracy is mediated by cognitive capacity.

4.1 Experiment 3a

To maximise our chances to replicate Alter et al.'s finding, we recruited students from a prestigious university, the Technion-Israel Institute of Technology. To maximize the similarity to Alter et al. (2007, Experiment 1) and to rule out potential concerns regarding media effects, this experiment was conducted on paper. We also collected Israeli SAT scores as a measure of IQ. Under the hypothesis that the effect of perceptual fluency is manifested only among reasoners of high ability, we predicted an interaction between SAT scores and the perceptual fluency manipulation, such that performance in the disfluent condition would be higher than the fluent condition only for high SAT scorers.

In addition to answers to the CRT, we also collected retrospective confidence from half the participants as an additional source of information about the subjective experience of our participants. Under the hypothesis that the fluent font should be processed easily and fluently and the disfluent font should create a sense of unease, we would expect confidence ratings to be higher for the fluent than the disfluent font condition.

4.1.1 Method

4.1.1.1 Participants

One hundred fifty undergraduate students of the Technion-Israel Institute of Technology volunteered to participate in the experiment. The mean age was 24.0 years (*SD* = 3.5); 28% were females. This was a high ability sample, as evidenced by their reported Israeli SAT scores, which ranged from 603 to 780 (*M* = 689, *SD* = 39[11]).

4.1.1.2 Materials.

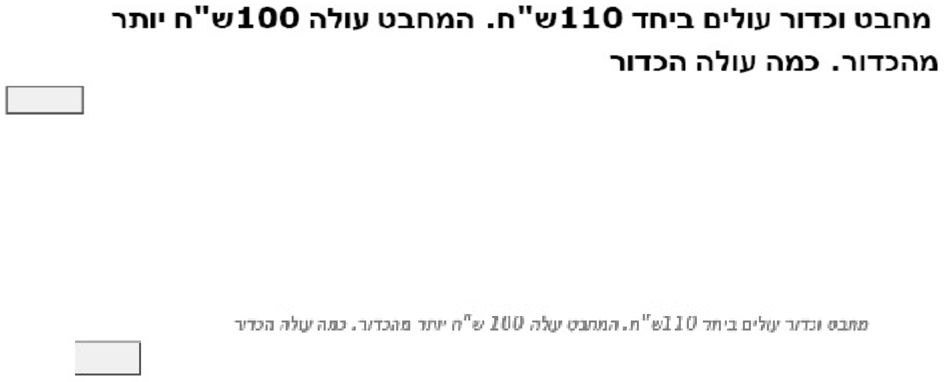
[11] Scores on this test range from 200 - 800; the national mean in 540; 15% nationally score above 650, 5% score above 700.

The three CRT problems used by Alter et al. (2007) were translated to Hebrew. A pre-test was used to choose the fluent and disfluent fonts for the study. Twenty participants rated the legibility of one base font and four other font types on a five-point scale ranging from 1 (illegible) to 5 (easier to read than the base font). The chosen fluent font was identified as easy to read (i.e., rated 4 or higher) by all participants. The chosen disfluent font was rated as illegible (1) by two participants, as legible with effort (2) by fourteen participants, and as legible but cause feeling of discomfort (3) by four participants. No participant characterized this font as easy to read (4) or easier than the regular font (5). Figure 1 presents the fluent and disfluent versions of the bat and ball problem.

The three CRT problems were printed on one white page in the order they appeared in Alter et al. (2007). At the bottom of the page there were spaces for writing demographic information. There were four versions for the printed page: Two with each font type, with each one of them either including a confidence rating scale after each problem, or not. When included, confidence was rated by choosing a number on an 11-point scale marked by 0, 10,… 100%.

4.1.1.3 Procedure.

The participants were recruited at the campus centre (see Alter et al., 2007), grass plots, libraries, and faculty lobbies all around the campus. Each participant randomly received one of the four questionnaire versions (*N* > 30 in each group). They were asked to solve each problem and to indicate whether they were familiar with the problem before taking the test. Participants who knew at least one problem in advance were replaced.



מחבט וכדור עולים ביחד 110ש"ח. המחבט עולה 100ש"ח יותר
מהכדור. כמה עולה הכדור

מחבט וכדור עולים ביחד 110ש"ח. המחבט עולה 100 ש"ח יותר מהכדור. כמה עולה הכדור

Figure 1. Experiment 3a: Hebrew versions of the bat and ball problem presented in difficult and easy to read fonts.

4.1.2 Results.

To establish that the four groups of participants were equivalent, we analysed SAT scores, age, and Hebrew proficiency with a 2 (perceptual fluency) x 2 (confidence elicitation) between-subjects ANOVA. None of the main effects or interactions was significant. Providing confidence ratings was not reactive with respect to CRT performance, $t < 1$ for accuracy difference between the groups. Thus, except for the reporting of the results of confidence ratings, all data analyses were collapsed across the groups who rated or did not rate their confidence in each answer.

Overall success rate was (2.0/3) similar to the success rates found in Alter et al.'s (2007) sample. The confidence ratings did not suggest that the difficult to read font lowered confidence ratings: For the participants who rated confidence in each answer ($N = 77$), confidence was equivalent for the easy to read ($M = 90\%$; $SD = 12$) and the difficult to read fonts ($M = 91\%$, $SD = 10$), $t(75) < 1$.

The following analyses test the hypothesis that the effects of the perceptual fluency manipulation would vary as a function of SAT scores. Overall, as found in previous studies (Frederick, 2005), the success rate was correlated with SAT scores ($r = .34$, $p < .01$). To analyse the interaction between SAT scores and the perceptual fluency manipulation, an ANOVA was performed with font type as a dichotomous variable and SAT score as a random factor. No main effects were found, both Fs < 1, but there was a significant interaction, $F(22,42) = 1.89$, $MSE = 763.46$, $p < .05$, $\eta_p^2 = .50$.

To understand the nature of the interaction, we divided the sample into four quartiles based on SAT scores and plotted percentage correct as a function of the font type. These data are presented in Figure 2. As is clear from the figure, the interaction is driven by the medium-high-SAT group, for whom performance was significantly lower in the fluent font than the disfluent font, $t(38) = 3.86$, $p < .0001$, although there was no difference in the confidence ratings for the two conditions ($M = 94\%$, $SD = 8$). In all other quartiles, the success rate and confidence ratings were equivalent for the two font types, all *t*s < 1.

4.1.3 Conclusions

Despite having been sampled from a prestigious university, the effect of the perceptual fluency manipulation over the entire sample was small. Thus, our earlier failure to replicate cannot be completely attributed to either differences in participant populations. It is clear, however, that performance differences between font types varied as a function of cognitive ability. In this sample, the difference emerged in the 3rd quartile of participants, consistent with the hypothesis that the manipulation is effective only among the high ability participants. The effect of the disfluent font condition was to bring performance for the 3rd quartile participants level to the 4th quartile participants. It is likely that no further effect was observed amongst the top participants because their performance in the fluent condition was almost at ceiling. Consistent with our earlier findings that FOR judgments did not vary as a function of perceptual fluency, in the current study we observed that retrospective confidence judgments were similar between the two groups.

4.2 Experiment 3b

The goal of this final study was twofold. First, we wanted to replicate the finding that the Type 2 thinking induced by perceptual disfluency advantages only the more cognitively able reasoners. Our second goal was to rule out an alternative reason for the observed relationship

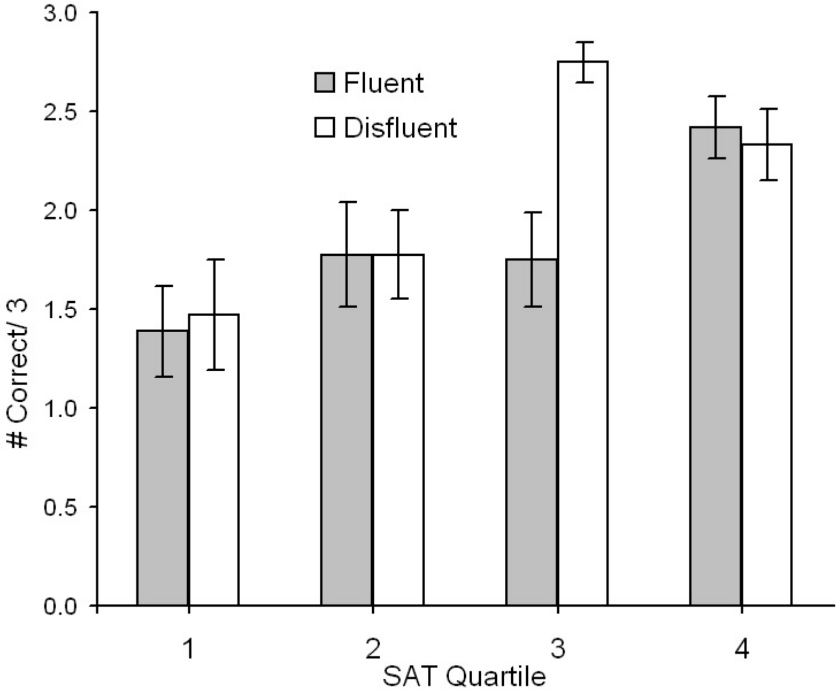


Figure 2. Experiment 3a: CRT scores for fluent and disfluent fonts as a function of self-reported SAT scores.

between cognitive ability and the fluency manipulation, namely a disposition to engage in effortful thinking, which has been shown to be related to normatively correct performance on a variety of reasoning tasks (Stanovich, 1999). To do this, we included a self-report measure of thinking dispositions called the Actively Open-minded Thinking scale (Stanovich & West 2007; 2008). We also collected retrospective confidence judgments and measured RT for each trial.

4.2.1 Method

4.2.1.1 Participants

A total of 142 people participated either from the University of Saskatchewan introductory psychology participant pool or from advertisements posted around the campus; 68% were female with a mean age of 20.8 years (*sd* = 4.1). About half (54%) reported high school as their highest level of education, 24% had completed a Bachelor's degree, 19% had completed at least one year of undergraduate studies, and the remainder had a college diploma.

4.2.1.2 Procedure

Participants were tested individually and were randomly assigned to the difficult and easy to read font conditions. The CRT problems were presented on a computer in a 10 point Courier New black font on a white background (easy) or a teal italicized 10 point Curlz MT font on a green background (difficult, as described in Experiment 1a). After completing each problem, participants were asked to rate their confidence on a 7-point scale with "7" representing the highest level of confidence.

After completing the CRT, participants were administered the Shipley Institute of Living Scale, which was used to derive estimates of IQ. They also completed the Actively Open-minded Thinking Scale (AOT; Stanovich & West 2007; 2008); this is a 41-item self-report measure of an inclination to engage in effortful vs intuitive thinking (e.g., "No one can talk me out of something I know is right" and "If I think longer about a problem I will be more likely to solve it"). Participants respond on a six point scale; high scores indicate a preference for actively open-minded thinking[12]. The time required to complete the experiment was about 30 minutes.

4.2.3 Results

As before, RT's were converted to $\log^{10}$ prior to analysis. IQ scores were equivalent for those in the difficult (M =113, SD = 7.2) and easy to read font conditions (M = 111, SD = 10.8), $t(140) = 1.24$, $p = .22$, as were AOT scores, $t < 1$ (M = 180, SD = 16.3 and M = 177, SD = 17.9, respectively). Performance on the CRT was correlated with both IQ, $r = .43$, $p < .001$ and AOT scores, $r = .32$, $p < .001$; these latter two were also correlated, $r = .29$, $p < .001$. As was the case in our earlier studies, there was a tendency for reasoners to take longer for the difficult (M = 36.5 sec, SD = 19.2) than the easy to read font (M = 31.0 sec, SD = 15.7), $t(140) = 1.88$, $p = .06$, d = .32. Nevertheless, this extra time did not facilitate CRT performance, which was similar for the two groups, $t < 1$ (M = 1.05/3, SD = .13 and M = 1.14/ 3, SD = .12 for the difficult and easy to read fonts respectively).

To test the hypothesis that the effectiveness of the fluency manipulation varies as a function of IQ, we computed an analysis of covariance (ANCOVA) with AOT and IQ entered as continuous variables, fluency condition as a dichotomous between-subjects variable, as well as the IQ by fluency interaction. All of the main effects were significant, $F(1, 137) \geq 5.09$, MSE = .11, $p \leq .026$, $\eta_p^2 \geq .04$ as was the predicted interaction between IQ and fluency condition, $F(1, 137) = 4.65$, $p = .033$, $\eta_p^2 = .03$. These findings replicate Experiment 3a, supporting the

[12] Toplak, West and Stanovich (2011) report Cronbach's alpha for the scale to be .81

hypothesis that the effectiveness of the fluency manipulation depends on cognitive ability. Moreover, the relationship between IQ and fluency condition could not be attributed to differences in predisposition to open-minded thinking, as the interaction remained significant after partialling out variance attributed to AOT scores.

To understand the nature of the interaction, we divided the sample into four quartiles based on IQ scores. These data are presented in Figure 3. These data show a clear cross-over pattern, with performance in the low IQ group better on the easy than difficult to read font, while the reverse was true for the high IQ group, although neither comparison was significant, $t(35) = 1.23$, $p = .22$ for the lowest quartile and $t(30) = 1.14$, $p = .26$ for the highest quartile. Nonetheless, the fact that the interaction was significant clearly supports the hypothesis that the effect of the fluency manipulation varies as a function of the cognitive capacity of the reasoner.

Finally, as was the case in Experiment 3a, there was no evidence that the perceptual fluency manipulation affected confidence judgments, $t < 1$. As expected, however, confidence was inversely related to RT, $r = -.33$, $p <.001$, a relationship that was observed in both the fluent and disfluent font conditions, $r = -.36$, $p = .001$ and $-.33$, $p = .002$. Thus, as was the case with FOR judgments, retrospective confidence judgments were higher for answers that were produced quickly than those produced slowly.

4.2.4 Conclusions

As in our previous studies, there was no overall effect of the perceptual fluency manipulation on correct responses, nor did confidence vary between the two conditions, contrary to our hypothesis. However, as was the case in our earlier studies, we found that perceptual

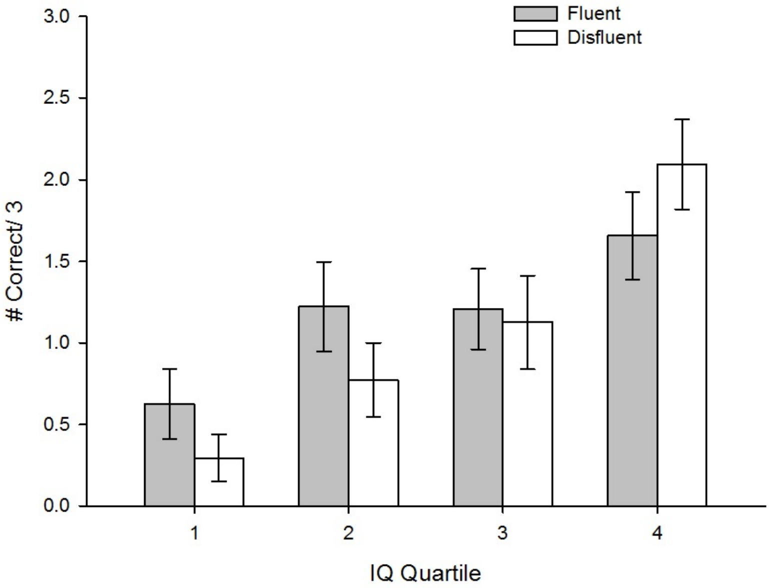


Figure 3. CRT scores for fluent and disfluent fonts as a function of IQ in Experiment 3b.

disfluency has the effect of slowing response times, suggesting that more Type 2 thinking may be taking place in the disfluent relative to the fluent condition. However, that additional processing time did not translate into higher accuracy, except for a tendency among the most cognitively able participants of the present sample, as was observed in Experiment 3a. These data suggest that relationship between perceptual fluency and accuracy was absent in our earlier studies because the high and low IQ participants contribute differently to the overall pattern.

5. General Discussion

Historically, the metacognitive issues of monitoring answers and the control of mental effort have been almost neglected in the reasoning literature. Whereas numerous studies have investigated the processes by which conclusions are generated, we know relatively little about the processes that determine when more or less effort is engaged to reach that answer. Until recently, when the issue of monitoring and control has been discussed, it has largely been assumed that Type 2 processes are responsible for both monitoring the output of Type 1 processes and of performing rule-based, Type 2 processing (see Evans, 2009 for a discussion). The present paper is built upon the ideas expressed by Thompson (2009, 2010) that metacognition is a third type of processes, responsible for regulating reasoning processes, while Type 2 processes are responsible for the deliberate processing itself. We thus concur with recent theorists (Evans, 2009; Stanovich, 2009), who have argued the need for a third set of processes whose role is to monitor the outputs of Type 1 processes and to determine the extent to which Type 2 processes are engaged.

In keeping with the extensive literature on metacognitive judgements, we posited that an experiential cue, fluency, is used to monitor reasoning outputs in the same manner as it is used to monitor memory outputs (e.g. Kelley & Lindsay, 1993; Koriat, Ma'ayan & Nussinson, 2006). In this paper, we call attention to the difference between two types of fluency: Answer fluency and perceptual fluency. Experiments 1a and 1b replicated our earlier findings that answer fluency was a reliable predictor of FOR judgments (Thompson et al., 2011); Experiment 3b extended this finding to retrospective confidence judgments. Fluent processing was associated with higher FOR ratings, and those were then associated with reduced Type 2 engagement as measured in terms of rethinking times and answer change. Moreover, the effect of answer fluency was observed regardless of whether the stimuli were difficult or easy to perceive. These data support the conclusions that a) answer fluency is used to monitor and control reasoning behaviour in a manner similar to that observed in the memory literature and b) its role in monitoring and control is independent from the effect of perceptual fluency.

It is clear that the effect of perceptual fluency is more subtle. In four of five experiments where thinking time was measured, we observed that the disfluent condition promoted increased thinking time, which is evidence of additional Type 2 thinking. However, this additional time did not produce more changes of answer, as was the case for answer fluency, nor did it result in a higher probability of correct answers, as Alter et al. (2007) observed. Thus, whilst the perceptual fluency manipulation was sufficient to trigger additional though, it did not necessarily trigger an inclination to change the initial answer. We posit that the sense of unease created by perceptual disfluency was sufficient to alert reasoners that something might be awry, but for most, their response to this unease was to spend more time rationalizing or justifying their initial answer, rather than changing it.

In our final two experiments, we tested the possibility that the outcome of the perceptual fluency manipulation varies with cognitive ability. Indeed, the expected pattern on correct answers was revealed, but only for relatively high ability participants, even when the sample was taken from an elite university. Importantly, Alter et al.'s (2007) sample was drawn from elite universities. A similar effect has been observed with instructional manipulations: Instructions that emphasize the need to reason logically and put aside beliefs are more effective among high capacity than low capacity participants (Evans, Handley, Neilens, & Over, 2010).

It is very likely that the key to reconciling our data with the observations that perceptual fluency affects a wide range of judgments and decisions is the extent to which the effect of fluency on performance is mediated by other variables, such as working memory capacity, the ability to inhibit the initial answer, the ability to understand and apply normative answers, etc. That is, while perceptual disfluency may produce longer RTs, this additional reasoning time will translate into correct performance only for a select few on complex reasoning problems. However, the effect might be relatively more manifest with more open-ended judgments for which fluency effects are commonly observed (see Alter & Oppenheimer, 2009 for a review). Consistent with this hypothesis, Morsanyi and Handley (2012) found that although participants liked syllogisms less when they were presented in a difficult-to-read font, their reasoning performance was not better in the disfluent condition.

Another way to view these data is that the source of the perceptual fluency effect might be more diffuse than the effect of answer fluency. That is, while answer fluency is a variable that can be monitored trial by trial, and thus serve as the input for a judgment of relative item difficulty, perceptual fluency varies across a set of items and may serve as a cue to their collective difficulty. Thus, perceptual disfluency may trigger a global sense of unease, which is sufficient to trigger longer thinking times, the outcome of which varies with the specific ability of the reasoner and, possibly, the difficulty of the task that is being reasoned about.

Another implication of this and of our previous work (Shynkaruk & Thompson, 2006; Thompson et al., 2011) is that measures of normative accuracy are not good measures of Type 2 engagement. We have found that longer thinking times are correlated with the probability of changing answers, but this does not necessarily guarantee accurate responding (Thompson et al., 2011); indeed, as in Experiment 1a, reasoners are often just as likely to change a correct answer to an incorrect one as vice versa (Pennycook & Thompson, 2012; Shynkaruk & Thompson, 2006). In many instances, reasoners do not change their initial, intuitive answer, suggesting that they are engaged in "serial associative cognition with a focal bias" (Stanovich, 2009, p. 67), perhaps in an attempt to rationalise or justify their initial choice (Evans, 1996; Evans & Ball, 2010; Stanovich, 2004; Wilson & Dunn, 2004). Thus, reasoners may engage in a large number of deliberate, analytic behaviours that do not result in a normatively correct answer (Stanovich, 2009). For these reasons, we concur with Elqayam and Evans (2011) that it is a fundamental error to equate normative accuracy and Type 2 thinking.

This may help to explain why the correlation between FOR and accuracy is either small (Thompson et al., 2011) or non-significant, as was the case in Experiment 1a. Low FORs were associated with longer thinking times, but if longer thinking times do not increase accuracy, then one would not necessarily expect a relationship between FOR and accuracy. Indeed, experiential cues, such as fluency, are diagnostic only insofar as they arise from cues that are correlated with accuracy (Ackerman & Zalmanov, in press; Begg, Duft, Lalonde Melnick, & Sanvito, 1989). In some domains, such as question answering, variability in response time for a given participant reliably differentiates right and wrong answers, such that answers that are provided quickly are more often correct than answers provided after a longer delay (e.g., Robinson, Johnson, & Herndon, 1997). In contrast, if a fluently generated answer is misleading, then the fluency cue is a pitfall (Ackerman & Zalmanov, in press; Benjamin et al., 1998; Koriat, 2008). Indeed, the reason that so many of the classic reasoning problems produce confidently held errors may be that the initial answer is generated with compelling ease (Simmons & Nelson, 2006; Thompson et al., 2011; see Kahneman, 2003 and Stanovich, 2004 for a discussion of the origins of intuitive responses).

Finally, our data suggest that the relation between fluency and Type 2 behaviour is complex. For example, cognitive ability appears to moderate the relationship between perceptual fluency and normative accuracy. A variety of other factors may be expected to moderate this relationship, including the perceived relevance of the feeling of fluency to the task at hand (Greifeneder et al., 2011), the motivation of the reasoner (Greifeneder et al., 2011; Stanovich, 2009), and the extent to which the allocation of resources is under participant control (Tullis & Benjamin, 2011). Thus, while we have sketched the outline of a metacognitive theory of reasoning, more research is needed to untangle the relationships between fluency, FOR, and Type 2 thinking.

References

Ackerman, R., & Goldsmith, M. (2008). Control over grain size in memory reporting — with and without satisficing knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 1224-1245.

Ackerman, R., & Zalmanov, H. (In press). The persistence of the fluency-confidence association in problem solving. *Psychonomic Bulletin & Review.*

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219-235.

Alter, A. L.,Oppenheimer, D.M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*, 569-576.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610-632.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55-68.

Brase, G.L., Fiddick, L., & Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. *The Quarterly Journal of Experimental Psychology, 59 (5)*, 965-976.

Briñol, P., Petty, R. E., Tormala, Z. L. (2006). The malleable meaning of subjective ease. *Psychological Science, 17*, 200-206.

Cokely, E.T., Parpart, P., & Schooler, L.J. (2009). On the link between cognitive control and heuristic processes. In N.A. Taatgnen & H. Van Rijn (Eds.), *Proceedings of the 31 Annual Conference of the Cognitive Science Society*, pp 2296-2931. Austen TX: The Cognitive Science Society.

Denes-Raj, V. & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology, 66*, 819-829.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual-process theories of thinking. *Cognition, 106*, 1248-1299.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12*, 83-87.

Elqayam, S. & Evans, J. St. B. T. (2011). Subtracting "ought" from "is": Descriptivism versus normativism in the study of human thinking. *Behavioral & Brain Sciences, 34(5),* 233-290.

Evans, J. St. B. T. (1996). Deciding before you think: Relevance and reasoning in the selection task. *British Journal of Psychology, 87*, 223-240.

Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review, 13*, 378-395.

Evans, J. St. B. T. (2007). *Hypothetical Thinking: Dual-processes in Reasoning and Judgment*. New York: Psychology Press.

Evans, J. St. B. T. (2009). How many dual-process theories do we need: One, two, or many? In J. Evans and K. Frankish (Eds.) *In Two Minds: Dual Processes and Beyond*, pp 33-54. Oxford University Press.

Evans, J. St. B. T. & Ball, L.J. (2010). Do people reason on the Wason selection task? A new look at the data of Ball et al. (2003). *The Quarterly Journal of Experimental Psychology, 63*, 434-441.

Evans, J. St. B. T., Barston, J. L. & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition , 11*, 295-306.

Evans, J. St. B. T., Handley, S. J., & Harper, N. J. (2001). Necessity, possibility and belief: A study of syllogistic reasoning. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 54*, 935-958.

Evans, J. St. B. T., Handley, Simon J., Neilens, H., & Over, D. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *The Quarterly Journal of Experimental Psychology, 63*, 892-909.

Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000) The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making, 13*, 1-17.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19,* 25-42.

Gigerenzer, G., Todd, P., & ABC Research Group. (1999). *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Greifeneder, R., Bless, H., & Pham, M. T. (2011). When do people rely on affective and cognitive feelings in judgement? A review. *Personality and Social Psychology, 15(2)*, 107-141.

Hertwig, R. Herzog, S.M., Schooler, L.J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental*

Psychology: Learning, Memory, & Cognition, 34, 1191-1206.

Hogarth, R. (2010). Intuition: A challenge for psychological research on decision making. *Psychological Inquiry, 21*, 338-353.

Kahneman, D. (2000). A psychological point of view: Violations of rational rules as a diagnostic of mental processes. *Behavioral and Brain Sciences, 23*, 681-683.Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. American Psychologist, 58, 697-720.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58 ,* 697-720.

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

Kelley, C.M., & Lindsay, S.D. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language, 32*, 1-24.

Klauer, K. C., Musch, J. & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review, 107,* 852-884.

Koriat, A. (2007). Metacognition and consciousness. In P.D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge Handbook of Consciousness*, (pp 289-326). Cambridge, NY: Cambridge University Press.

Koriat, A. (2008). Subjective confidence in one's answers: the consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34(4)*, 945-959.

Koriat, A., & Ackerman, R. (2010). Choice latency as a cue for children’s subjective confidence in the correctness of their answers. *Developmental Science*, *13*, 441-453.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*, 643 - 656.

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*, 478-492.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General, 135*, 36-69.

Koriat, A., & Nussinson, R. (2009). Attributing study effort to data-driven and goal-driven effects. Journal of Experimental Psychology: Learning, Memory and Cognition, 35, 1338-1343.

Metcalfe, J., & Finn, B. (2008). Evidence that judgements of learning are causally related to study choice. *Psychonomic Bulletin & Review*, 15, 174-179.

Morsanyi, K., & Handley, S. J. (2012). Logic Feels So Good—I Like It! Evidence for Intuitive Detection of Logicality in Syllogistic Reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(3)* 596-616. doi:10.1037/a0026099

Oaksford, M. & Chater, N. (2007). *Bayesian Rationality*. Oxford: Oxford University Press.

Pennycook, G., Fugelsang, J.A. & Koehler, D.J. (2012). Are we good at detecting conflict during reasoning? *Cognition, 124,* 101-106.

Pennycook, G & Thompson, V.A. (2012). Reasoning with base rates is routine, relatively effortless, and context-dependent. *Psychnonomic Bulletin and Review*, *19(3),* 528-534. doi: 10.3758/s13423-012-0249-3.

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

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition, 8*, 338-342.

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *Journal of Applied Psychology, 82*, 416-425.

Shynkaruk, J. M., & Thompson, V. A. (2006). Confidence and accuracy in deductive reasoning. *Memory & Cognition, 34,* 619-632.

Simmons, J. P., & Nelson, L. D., (2006). Intuitive confidence: Choosing between intuitive and nonintuitive alternatives. *Journal of Experimental Psychology: General, 135*, 409-428.

Sinclair, M. (2010). Misconceptions about intuition. *Psychological Inquiry, 21,* 378-386.

Singer, M., & Tiede, H. L. (2008). Feeling of knowing and duration of unsuccessful memory search. *Memory & Cognition, 36*, 588-597.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin,119*, 3 – 22.

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 3,* 601-604.

Stanovich, K. E. (1999). *Who is Rational?: Studies of Individual Differences in Reasoning.*

Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Stanovich, K. E. (2004). *The Robot's Rebellion: Finding Meaning in the Age of Darwin.* Chicago, ILL: The University of Chicago Press.

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Eds.), *In Two Minds: Dual Processes and Beyond* (pp. 55-88). Oxford: Oxford University Press.

Stanovich, K. E., & West, R. F. (2007). Natural my side bias is independent of cognitive ability. Thinking & Reasoning, 13, 225-247.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. Journal of Personality and Social Psychology, 94, 672-695.

Stepper, S., & Strack, F. (1993). Proprioceptive determinants of emotional and nonemotional feelings. Journal of Personality and Social Psychology, 64, 211-220.

Stupple, E.J.N., & Ball, L.J. (2008). Belief-conflict resolution in syllogistic reasoning: Inspection time evidence for a parallel process model. *Thinking & Reasoning, 14*, 168-181.

Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory and Cognition, 22,* 742-758.

Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. Evans and K. Frankish (Eds.) *In Two Minds: Dual Processes and Beyond*. Oxford: Oxford University Press.

Thompson, V. A. (2010). Towards a dual-process model of conditional inference. In M. Oaksford (Ed.) *The Psychology of Conditionals*. Oxford: Oxford University Press.

Thompson, V.A., ProwseTurner, J., & Pennycook, G. (2011). Intuition, Metacognition, and Reason. *Cognitive Psychology, 63,* 107-140.

Thompson, V. A., Striemer, C. L., Reikoff, R., Gunter, R. W., Campbell, J. I. D. (2003). Syllogistic reasoning time: Disconfirmation disconfirmed. *Psychonomic Bulletic & Review, 10*, 184-189.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics and biases tasks. *Memory & Cognition, 39*, 1275-1289.

Topolinski,S., & Reber, R. (2010). Gaining insight into the "aha" experience. *Current Directions in Psychological Science*, *19*(6), 402-405.

Tullis, J.G. & Benjamin, A.S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64*, 109-118.

Wilson, T. D., & Dunn, E. W. (2004). Self-knowledge: Its limits, value, and potential for improvement. *Annual Review of Psychology, 55*, 493-518.

Appendix

CRT Problems (correct responses in parentheses):

1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

   ____ cents **(5)**

2) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

   ____ days **(47)**

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

   ____ minutes **(5)**

Author Note

We would like to thank Melanie Pritchford for her help programming Experiments 2a and 2b and Sean Sacher for his assistance with editing the manuscript. Funding for Experiments 1a, 1b, 2c, and 3b was provided by the Natural Sciences and Engineering Research Council of Canada. Funding for Experiment 3a was provided by a grant from the Israel Foundation Trustees (2011-2013). Correspondence regarding this manuscript should be addressed to Valerie Thompson, Department of Psychology, University of Saskatchewan, 9 Campus Drive, Saskatoon SK, Canada S7N 5A5 or by email: valerie.thompson@usask.ca