Cognitive Psychology 80 (2015) 34–72

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

What makes us think? A three-stage dual-process model of analytic engagement

Gordon Pennycook *, Jonathan A. Fugelsang, Derek J. Koehler

Department of Psychology, University of Waterloo, Canada

ARTICLE INFO

Article history:
Accepted 15 May 2015
Available online 16 June 2015

Keywords:
Dual-process theory
Conflict detection
Conflict monitoring
Biases
Reasoning
Decision making
Base-rate neglect

ABSTRACT

The distinction between intuitive and analytic thinking is common in psychology. However, while often being quite clear on the characteristics of the two processes ('Type 1' processes are fast, autonomous, intuitive, etc. and 'Type 2' processes are slow, deliberative, analytic, etc.), dual-process theorists have been heavily criticized for being unclear on the factors that determine when an individual will think analytically or rely on their intuition. We address this issue by introducing a three-stage model that elucidates the bottom-up factors that cause individuals to engage Type 2 processing. According to the model, multiple Type 1 processes may be cued by a stimulus (Stage 1), leading to the potential for conflict detection (Stage 2). If successful, conflict detection leads to Type 2 processing (Stage 3), which may take the form of rationalization (i.e., the Type 1 output is verified *post hoc*) or decoupling (i.e., the Type 1 output is falsified). We tested key aspects of the model using a novel base-rate task where stereotypes and base-rate probabilities cued the same (non-conflict problems) or different (conflict problems) responses about group membership. Our results support two key predictions derived from the model: (1) conflict detection and decoupling are dissociable *sources* of Type 2 processing and (2) conflict detection sometimes fails. We argue that considering the potential *stages* of reasoning allows us to distinguish early (conflict detection) and late (decoupling) sources of analytic thought. Errors may occur at both stages and, as a consequence, bias arises from *both* conflict monitoring and decoupling failures.

* Corresponding author at: Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada.

E-mail address: gpennyco@uwaterloo.ca (G. Pennycook).

http://dx.doi.org/10.1016/j.cogpsych.2015.05.001
0010-0285/

truthiness (noun)

1: "truth that comes from the gut, not books" (Stephen Colbert, Comedy Central's "The Colbert Report," October 2005).

1. Introduction

A few months after the 2003 invasion of Iraq, current Vice President and then Senator Joe Biden asked President George W. Bush how he can be so sure that the United States was on the right course. Bush responded by putting his hand on the Senator's shoulder and saying "my instincts" (Suskind, 2004). Bush's faith in his gut feelings in the face of conflicting or contradictory evidence is, not incidentally, reminiscent of comedian Stephen Colbert's concept of "truthiness".[1] There appears to be a great deal of truth to the idea of truthiness and, in fact, it has been known for decades, dating back to Kahneman and Tversky's heuristics and biases research program, that humans often rely on intuition when making decisions (Tversky & Kahneman, 1974; for a recent overview, see Kahneman, 2011).

An additional point that is rarely emphasized, however, is that gut feelings do not *always* predominate. Some individuals are less likely to "go with their gut" when reasoning (Stanovich & West, 1998, 2000) and problems that cue conflicting response outputs have been shown to lead to deliberative reasoning (De Neys & Glumicic, 2008; De Neys, Vartanian, & Goel, 2008). Investigations of the factors that *undermine* intuitive decision making may lead to interventions which could be used to avoid future errors; or, in other words, to maximize "truth" and minimize "truthiness". To that end, it has been suggested that one of psychological science's most pressing goals should be to "give debiasing away" to the general public (Lilienfeld, Ammirati, & Landfield, 2009).

We argue that basic cognitive research that elucidates how debiasing happens in the absence of explicit top-down intervention could be a fruitful source of practical benefit in the public sphere. In the current work, we attempt to elucidate the cognitive processes that guard against reasoning failures by introducing a three-stage dual-process model of analytic engagement, along with 4 experiments that test predictions generated from the model. Our goal is to integrate perspectives on bias and irrationality that have previously been considered antithetical by breaking the reasoning process into stages and components. We argue that a consideration of the bottom-up *sources* of analytic thinking offers a new perspective that leads to novel predictions.

1.1. Dual-processing

Human reasoning and decision-making is thought to involve two distinct types of processes (for reviews, see Evans, 2008, 2010a; Evans & Stanovich, 2013a; Frankish & Evans, 2009; Sloman, 1996; Stanovich, 2004): Type 1 processes that are intuitive, fast, autonomous, and high capacity; and Type 2 processes that are reflective, slow, and resource demanding. Type 1 processes are thought to provide default outputs that may be acted upon as explicit representations manipulated in working memory via Type 2 processing (Evans & Stanovich, 2013a; Thompson, 2013). However, the question of what leads someone to engage deliberate and effortful reasoning in lieu of more intuitive and automatic cognitive processes is still unclear and, as a result, has been the focus of much recent scholarship

[1] The following is the full quote from Suskind's *New York Times* article: "Forty democratic senators were gathered for a lunch in March just off the Senate floor. I was there as a guest speaker. Joe Biden was telling a story, a story about the president. "I was in the Oval Office a few months after we swept into Baghdad," he began, "and I was telling the president of my many concerns" – concerns about growing problems winning the peace, the explosive mix of Shiite and Sunni, the disbanding of the Iraqi Army and problems securing the oil fields. Bush, Biden recalled, just looked at him, unflappably sure that the United States was on the right course and that all was well. "'Mr. President,' I finally said, 'How can you be so sure when you know you don't know the facts?'" Biden said that Bush stood up and put his hand on the senator's shoulder. "My instincts," he said. "My instincts." Biden paused and shook his head, recalling it all as the room grew quiet. "I said, 'Mr. President, your instincts aren't good enough!'" The democrat Biden and the Republican Bartlett are trying to make sense of the same thing – a president who has been an extraordinary blend of forcefulness and inscrutability, opacity and action."

and empirical research (e.g., De Neys & Bonnefon, 2013; De Neys & Glumicic, 2008; Evans, 2009; Stanovich, 2009a; Thompson, 2009; Thompson, Prowse Turner, & Pennycook, 2011).

One of the criticisms of dual-process theories is that they describe the characteristics of the two processes but are unclear on the question of how they operate (De Neys & Glumicic, 2008; Evans, 2007, 2010b; Gigerenzer & Regier, 1996; Osman, 2004; Stanovich & West, 2000). A common claim among dual-process theorists is that Type 2 processes monitor the output of Type 1 processes (e.g., Evans, 2006; Kahneman & Frederick, 2005; Stanovich, 1999). This default-interventionist perspective explains how Type 2 processing can be biased by earlier Type 1 outputs. However, the idea that Type 2 processes are themselves responsible for the instantiation of Type 2 processing is clearly problematic. In contrast, parallel form dual-process theories posit that both types of processing operate in parallel from the outset of reasoning (e.g., Sloman, 1996, 2002; Smith & DeCoster, 2000). These parallel-competitive models suggest that bias is common because fast Type 1 processes output a response *before* the slower Type 2 processes can complete, though additional Type 2 processing may occur if the two types of processing output conflicting responses (for a comparison of default-interventionist and parallel-competitive models, see Evans, 2007; Evans & Stanovich, 2013a; Handley & Trippas, 2015). Parallel form theories highlight conflict detection as a source of *later* Type 2 processing, but still assume that the monitoring of conflict is itself a Type 2 process. Thus, as has been outlined elsewhere (Evans, 2009; Evans & Stanovich, 2013a), neither of the major groups of dual-process theories adequately explain important aspects of cognitive architecture because both assume that Type 2 processing is effectively caused by *itself*. This is a problem of particular importance because the utility and explanatory value of dual-process theories is thought to depend, at least partially, on our understanding of the *sources* of analytic reasoning (Evans, 2009; Stanovich, 2009a; Thompson, 2009).

In the current work we introduce a new perspective on the factors that lead to Type 2 engagement. Our goal is to investigate the bottom-up (i.e., stimulus-triggered) processes that lead to increases in deliberative thought, independent from top-down factors such as instructional manipulations (e.g., Evans, Handley, Neilens, Bacon, & Over, 2010) and individual differences in analytic thinking disposition (e.g., Stanovich & West, 2008). We combine insights from recent dual-process models (De Neys, 2012; Evans, 2009; Stanovich, 2009a; Thompson, 2009) into a three-stage model of analytic engagement. Using a version of a classic decision making task, we provide evidence for two core claims derived from the model: (1) The detection of conflicts between intuitive responses is a key determinant of analytic engagement, but sometimes fails, and (2) the deliberative override of an intuitive response in lieu of an alternative is a later source of Type 2 processing that is dissociable from earlier increases in Type 2 processing attributable to conflict detection. Following previous research, we posit that analytic thinking may take the form of either rationalization (i.e., bolstering or verifying an intuitive response) or decoupling (i.e., overriding or falsifying an intuitive response in lieu of an alternative). Moreover, we qualify our findings in meaningful ways with a top-down source of Type 2 processing: individual differences in analytic thinking disposition. Our results indicate that reasoning failures can emerge from two sources: (1) Failing to detect bias (leading to a failure to think analytically; e.g., Evans, 2007; Kahneman, 2003), or (2) successfully detecting bias (e.g., De Neys, 2012), but failing to use analytic thought to override the intuitive response.

1.2. Conflict monitoring and analytic thinking

Although research has shown that the degree of involvement of Type 2 processing can be affected by top-down factors such as instructions (e.g., Daniel & Klaczynski, 2006; Evans, Newstead, Allen, & Pollard, 1994; Vadeboncoeur & Markovits, 1999), the amount of time permitted to think (e.g., Evans & Curtis-Holmes, 2005; Finucane, Alhakami, Slovic, & Johnson, 2000), and individual differences in thinking disposition (e.g., Stanovich & West, 1998, 2000) isolating lower level cognitive processes that lead to Type 2 processing are more important for our emerging understanding of the dynamic relation between Type 1 and Type 2 processes in the mind. Bottom-up factors can be used to determine which type of processing will dominate. Consider the following base-rate problem (from De Neys & Glumicic, 2008, adapted from Kahneman & Tversky, 1973):

In a study 1000 people were tested. Among the participants there were 995 nurses and 5 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 most likely?

(a) Paul is a nurse.

(b) Paul is a doctor.

This problem includes two pieces of information that point to alternative responses. The base-rate probability (i.e., 995 nurses versus 5 doctors) indicates that there is a 99.5% chance that Paul is a nurse. In contrast, the personality description contains stereotypes that are strongly diagnostic of a doctor. A great deal of research has demonstrated that participants tend to strongly favor the stereotypical information over the base-rate probability because the stereotype is the more intuitive source of information (see Barbey & Sloman, 2007 for a review). Thus, the base-rate problem is thought to engender an initial response based on the salient stereotypical information.

Recent research has also indicated that people are implicitly aware of the conflict between base-rate and stereotype, despite the apparent neglect or underweighting of the base-rates (De Neys, Cromheeke, & Osman, 2011; De Neys & Franssens, 2009; De Neys & Glumicic, 2008; De Neys et al., 2008), perhaps because extreme probabilities (as shown above) can be processed very rapidly (Pennycook, Fugelsang, & Koehler, 2012; Pennycook & Thompson, 2012; Pennycook, Trippas, Handley, & Thompson, 2014). Importantly for present purposes, one of the key pieces of evidence for the conflict detection hypothesis is an increase in response time (RT) for conflict (as above) versus non-conflict (e.g., if there were 5 nurses and 995 doctors above) base-rate problems *even when participants give the stereotypical response*.[2] Thus, detection of the conflict between base-rate and stereotype appears to lead to increased Type 2 processing (as reflected by increased RT) even in cases where participants give the response that is more intuitively salient. In support of this claim, De Neys et al. (2008) found increased activation in the anterior cingulate cortex, a region of the brain associated with conflict detection (see Bush, Luu, & Posner, 2000 for review), for stereotypical responses to incongruent problems relative to congruent problems. Given the fact that participants gave the stereotypical response despite the apparent increase in Type 2 processing, it is likely that they spent their time rationalizing the stereotype or, at the very least, weighing the stereotype against the base-rate probability (Pennycook & Thompson, 2012). This leads to the appearance of "effortful beliefs": i.e., belief processing that is analytic (Handley, Newstead, & Trippas, 2011; Handley & Trippas, 2015; Trippas, Verde, & Handley, 2014).

The central role of conflict detection as an initiator of Type 2 processing is evidenced by a wide range of measures across numerous reasoning tasks (see De Neys, 2012 for a review). Indeed, response conflict has long been an important concept in reasoning and decision making research (Evans, Barston, & Pollard, 1983; Kahneman, 2000; Kahneman & Tversky, 1982; Wilkins, 1928) and much neuropsychological work supports the idea that "conflict" problems are processed differently than "non-conflict" problems (Banks & Hope, 2014; Goel & Dolan, 2003; Liang, Goel, Jia, & Li, 2014; Prado, Kaliuzhna, Cheylus, & Noveck, 2008; Prado & Noveck, 2007; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Stollstorff, Vartanian, & Goel, 2012). Nonetheless, conflict monitoring is not included as a component in most formal dual-process reasoning models (e.g., Evans, 2009; Stanovich, 2009a; Thompson, 2009, but see De Neys, 2012; Handley & Trippas, 2015), perhaps because monitoring has been considered a Type 2 process (and therefore not a separate component). Moreover, the primary dual-process model that *does* incorporate conflict monitoring – De Neys' (2012, 2014) logical intuition model – focuses entirely on the processes that lead to successful conflict detection and therefore does not incorporate differences in the *quality* of Type 2 processing.

The primary goal of the current work is to develop a dual-process model that includes both a conflict monitoring stage and a Type 2 processing stage that differentiates between different levels of analytic engagement. This model could then accommodate both major perspectives on the primary cause of biased responding: (1) A failure to engage Type 2 processing (e.g., Evans, 2007; Kahneman, 2003),

[2] Since the base-rate and stereotype point to the same response for non-conflict (i.e., congruent) problems, this comparison is isolated to cases where participants gave the base-rate/stereotype response. Naturally, this accounts for the vast majority of responses for congruent problems.

and (2) successfully engaging Type 2 processing following conflict detection, but failing to override the biased response (e.g., De Neys, 2012). These two sources of bias have often been discussed in the context of a debate about the modal biased reasoner (see De Neys, 2014 for a review) and, as such, we will also frame the perspectives as conflicting. However, this should not be taken to mean that authors such as Evans (2007) deny the existence of conflict detection (see Evans, 2009) or that authors such as De Neys (2012) deny the existence of analytic engagement failures (see De Neys, 2014). Our goal here is to assess the *models* of bias by the respective authors, which include predictions for one or the other source of bias but that do not necessarily preclude other factors.

1.3. Cognitive decoupling and analytic thinking

Conflict monitoring is not the only bottom-up source of analytic thinking. For example, De Neys and Glumicic (2008) also reported an increase in RT for incongruent (i.e., conflict) problems relative to congruent (i.e., non-conflict) when participants gave the *base-rate* response to the incongruent problems. In this case, the apparent increase in Type 2 processing is potentially a result of a *rethinking* or *decoupling* process. Indeed, De Neys and colleagues have postulated that participants engaged additional resources to inhibit the prepotent stereotypical response (De Neys & Franssens, 2009; De Neys et al., 2008; Franssens & De Neys, 2009). That is, given the idea that stereotypes cue intuitive (Type 1) responses, additional Type 2 processing is therefore thought to be necessary to suppress and override the stereotype response in lieu of the base-rate response (Barbey & Sloman, 2007). Again, in support of this claim, De Neys et al. (2008) found increased activation in the right lateral prefrontal cortex (RLPFC) for base-rate responses to incongruent problems relative to congruent. The RLPFC is considered a key area involved in response inhibition (see Aron, Robbins, & Poldrack, 2004, for a review). Base-rate responses, like stereotypical responses, were associated with increased ACC activation. This indicates that participants were able to detect the conflict between base-rate and stereotypes for incongruent problems regardless of their ultimate response, but cases where the base-rate response was given involved an *additional* deliberative reasoning process relative to when the stereotypical response was given. Considering the association between the RLPFC and response inhibition along with the presumed intuitiveness of stereotypical information, it is plausible that this additional process consisted of participants actively suppressing the stereotypical response. In other words, cognitive decoupling appears to be a *later* source of Type 2 processing relative to conflict detection.

An additional point needs to be clarified. The claim that base-rate responses are usually accompanied by an active suppression of the salient stereotypical response via Type 2 processing is not the same as claiming that the base-rate response *necessarily* requires Type 2 processing to enter into reasoning (De Neys, 2007). Indeed, a recent set of experiments using an instruction manipulation illustrated that both base-rates and stereotypes appear to interfere with each other (Pennycook, Trippas, et al., 2014). This cross-interference was evident even when participants were forced to respond within a short time-deadline. This finding indicates that both base-rates and stereotypes cue Type 1 outputs (see also Brenner, Griffin, & Koehler, 2012). Under this account, stereotypes typically dominate reasoning because they cue intuitive responses that come to mind more quickly and fluently than the base-rates (Pennycook, Trippas, et al., 2014). Stereotypes, in other words, are a more salient source of information than base-rates, but both may cue Type 1 outputs. Moreover, decoupling should occur in cases when the base-rate response is provided because an intuitive response based on the stereotypical information is thought to have come first in the reasoning process and therefore needs to be overridden for an alternative response to be given.

1.4. A three-stage model of analytic engagement

Fig. 1 represents our theoretical position. The model was built to describe the reasoning process for a problem or cue that elicits multiple conflicting outputs. It formalizes and combines distinctions made by previous theorists (e.g., De Neys, 2012; Epstein, 1994; Evans, 2006; Evans & Stanovich, 2013a; Handley & Trippas, 2015; Sloman, 1996, 2014; Smith & DeCoster, 2000; Stanovich, 2004; Strack & Deustch, 2004; Thompson, 2009) by dividing an individual reasoning event into stages and components. In the first stage, autonomous Type 1 processes generate so-called "intuitive" responses.

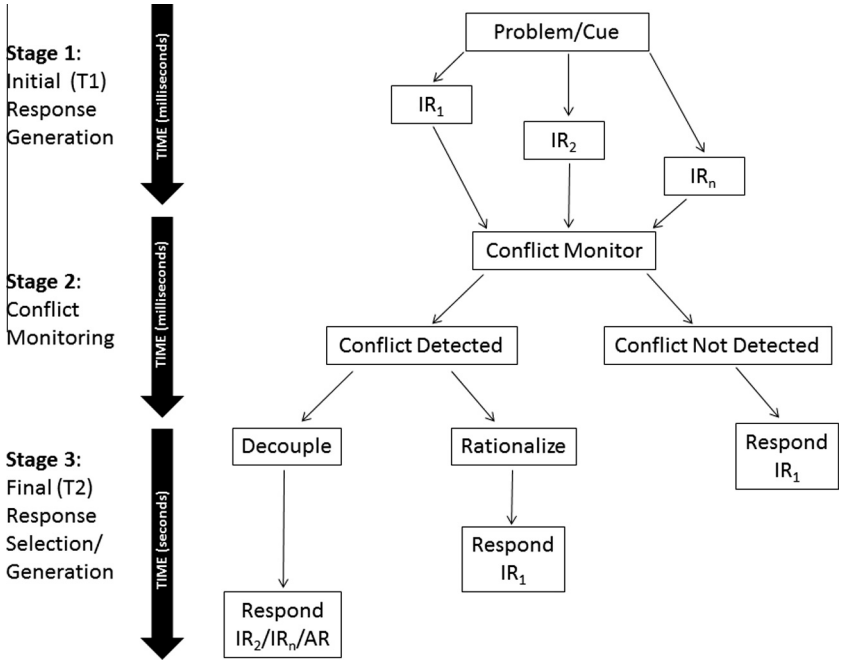

Fig. 1. Three-stage dual-process model of analytic engagement. T1 = Type 1 "intuitive" processing. T2 = Type 2 "analytic" processing. IR = initial response. IR's are numbered to reflect alternative speeds of generation. IR_1 is the most salient and fluent possible response. IR_n refers to the possibility of multiple, potentially competing, initial responses. AR = alternative response. IR_n refers to the possibility of an alternative response that is grounded in an initial response.

These Type 1 processes are cued by features of the stimulus, do not require working memory or executive functioning, and operate in parallel (Evans, 2008; Sloman, 1996; Stanovich, 2004). Given these features, we have inferred that some stimuli will cue multiple, potentially competing Type 1 outputs (for a similar perspective, see De Neys, 2012, 2014).

A second dimension of the initial stage in our model relates to the idea that some initial responses come to mind more quickly and fluently than others (Thompson, 2009; Thompson et al., 2011, 2013).[3] In the case of base-rate problems, for example, stereotypes are often used as *intuitive lures* because of the phenomenology of their fluent generation. However, this does not rule out the possibility that alternative sources of information can cue an alternative Type 1 output in parallel. As discussed above, extreme base-rates presented in simple frequency formats influence response time, confidence, and probability estimates in ways diagnostic of Type 1 processing (Pennycook, Trippas, et al., 2014). Thus, base-rate problems serve as an example of a case where two competing sources of information embedded in a problem can elicit competing initial responses (see Section 6.5 for further examples). The base-rate problem example is particularly powerful given the presumed alternative time-course of the stereotype initial response (IR_1) and the base-rate initial response (IR_2). Specifically, stereotypes likely cue initial responses that come to mind more quickly (and, as a consequence, more fluently) than do base-rates. For other types of problems or cues, it is possible that multiple additional initial responses are elicited, hence IR_n (see Fig. 1).

The role of the second stage, then, is to monitor for conflict between Type 1 outputs (De Neys, 2012, 2014). If no conflict is detected (either because no conflict existed or because of a conflict detection failure), the first initial response (IR_1) will continue to the third stage where it is accepted with cursory

[3] The key component of Thompson's (2009) model, metacognitive 'feelings of rightness', have not been integrated in the version of the three-stage model presented in Fig. 1 because our data do not speak to metacognitive considerations. This is an important area for future research.

analytic (Type 2) analysis. This is the prototypical way in which bias is thought to arise: unimpeded and with little effort. If a conflict is successfully detected, however, more substantive Type 2 reasoning will be engaged. Thus, conflict detection is a bottom-up source of analytic engagement.

The three-stage model then distinguishes between two very different forms of Type 2 processing, each with different implications for the degree of bias ultimately displayed. Rationalization is a form of Type 2 processing where, despite successful conflict detection, the reasoner focuses on justifying or elaborating the first initial response (IR_1) without seriously considering the Type 1 output that was cued by the stimulus, but that did not come to mind as quickly and fluently (IR_2) as the first initial response (IR_1).[4] This leads to a response in line with what would typically be considered bias (i.e., one's strongest intuition, which will often be personally relevant), but that has been bolstered by analytic reasoning (an "effortful" belief-based response; see Handley & Trippas, 2015). This process is traditionally referred to as "rationalization" in the reasoning literature (e.g., Wason & Evans, 1975), to highlight the idea that the additional consideration is focused on verifying, and not falsifying, the Type 1 output. For example, participants typically spend much of their time looking at the card they ultimately select on the Wason card selection task, indicating that they are likely focused on rationalizing their default response (Ball, Lucas, Miles, & Gale, 2003; Evans, 1996). We note, in addition, that rationalization is tied to a substantial body of research on motivated reasoning (see Kunda, 1990). This research indicates that the instantiation of Type 2 reasoning can sometimes lead to the *strengthening* of a pre-existing belief or attitude, particularly if the belief or attitude is of some personal significance.

The second class of Type 2 processes that could result from conflict detection is cognitive decoupling (Stanovich, 2004, 2009a). This is perhaps the most prototypical "analytic" process and, as such, has dominated the literature on reasoning. Decoupling refers to the additional processing necessary to inhibit and override an intuitive response (primarily, IR_1). There are three obvious possibilities given a decoupling process: (1) IR_1 is suppressed in lieu of IR_2 which, upon reflection, emerges as a stronger alternative, (2) IR_1 is suppressed in lieu of some other initial response (IR_n), and (3) an alternative response (AR) is generated that represents a novel amalgamation of initial responses (see Section 6.7.3 for further comment on AR).

The three-stage model is a novel combination of multiple perspectives, which means that individual aspects of the model are grounded in previous theory. The idea that some intuitive responses come to mind faster than others is an aspect of Thompson's (e.g., 2009) metacognitive dual-process theory. The idea that conflicting Type 1 outputs may cue analytic thinking is the core of De Neys' (e.g., 2012) logical intuition model (see also Handley & Trippas, 2015). Rationalization (e.g., Wason & Evans, 1975) and decoupling (e.g., Stanovich, 1999) have long been discussed in the context of dual-process models, though to the best of our knowledge they have not been included as separate classes of Type 2 processing in the same model. Moreover, Stanovich and West (2008) have used stages to determine *when* and *if* intuitive response will be overridden under the goal of creating a framework for understanding individual differences in heuristics and biases tasks (see also Kahneman & Tversky, 1982). Here, in contrast, we consider the reasoning *process* as stages to highlight different bottom-up *sources* of Type 2 processing. Previous models have highlighted key aspects of the reasoning process, but have largely left unanswered the question of what actually causes Type 2 processing to be engaged.

The goal of the current work is to demonstrate the utility of our three-stage model. This will be done in three ways: (1) By investigating the possibility that conflict monitoring may sometimes fail (Stage 2), (2) by dissociating increases in Type 2 processing that indicate, on one hand, rationalization following successful conflict detection and, on the other hand, cognitive decoupling (Stage 3), and (3) by investigating the locus of individual differences in reasoning.

1.5. Individual differences in analytic thinking

Prior to outlining our specific predictions, it is necessary to discuss an additional source of Type 2 processing. Research has indicated that the mere *willingness* to engage deliberative reasoning (i.e., differences in thinking disposition or cognitive style) predicts reasoning performance over and above

[4] We note that it is possible for more than two Type 1 outputs to be cued by a stimulus. Here we isolate our discussion to just two for the sake of simplicity.

individual differences in the *ability* to think analytically (i.e., cognitive ability or intelligence) (for reviews, see Stanovich, 2004, 2009a, 2011; Stanovich & West, 2000). For example, individuals who are actively open-minded are more willing to question and perhaps rethink an initial response (Baron, 2008). This disposition, as assessed by a number of questionnaires, has been linked to a wide range of reasoning and decision-making tasks (Stanovich & West, 1997, 1998; Toplak, West, & Stanovich, 2011). Taking the base-rate problem as an example, participants who are actively open-minded are more likely to choose the base-rate over the stereotype relative to less analytic individuals, presumably because they were more willing to think analytically about the initial stereotypical response (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014a). Stanovich (e.g., 2004, 2009b) has argued that thinking disposition is an underappreciated determinant of psychological outcomes. Recent research has supported the idea that cognitive style plays a consequential role in psychological domains that are of some general import: e.g., creativity (Barr, Pennycook, Stolz, & Fugelsang, 2014), moral judgments and values (Paxton, Unger, & Greene, 2012; Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014b; Rozyman, Landy, & Goodwin, 2014), religious belief (Gervais & Norenzayan, 2012; Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012; Pennycook, Cheyne, Koehler, & Fugelsang, 2013; Pennycook et al., 2014a; Shenhav, Rand, & Greene, 2012), and even Smartphone technology use (Barr, Pennycook, Stolz, & Fugelsang, 2015). The research indicating that individual differences in cognitive style have important effects on beliefs and behavior implies that the engagement of Type 2 reasoning processes involves an important dispositional component. Cognitive style has particular relevance for the current discussion as it represents an independent top-down source of Type 2 processing. That is, how much someone values or enjoys analytic thinking may contribute to the probability that they engage Type 2 process, independent of any Type 1 output monitoring process and therefore regardless of the content of the stimulus.

The foregoing highlights an additional source of uncertainty about the factors that elicit Type 2 processing; namely, do individual differences relate to conflict detection? Recently, De Neys and Bonnefon (2013) theoretically integrated research on conflict detection with individual differences in reasoning. Specifically, they asked the question "do biased and unbiased reasoners take different paths early on in the reasoning process or is the observed variance late to arise?" (p. 172). The answer to this question has significant implications: If individual differences only affect reasoning at a relatively late stage (Stage 3 in our model), as De Neys and Bonnefon claim, it would imply that the influence of said individual differences has been greatly overemphasised in the reasoning and decision making literature. To support this argument, De Neys and Bonnefon cited the many cases where even "biased" reasoners appeared to have detected reasoning conflicts, with respect to both RT increases for incongruent base-rate problems (De Neys & Glumicic, 2008), and many other types of problems and measures (De Neys, 2012). These findings suggest that "biased" and "unbiased" reasoners diverge late in the reasoning process, thereby suggesting that both types of reasoners are likely closer in cognitive function to each other than some may have previously considered.[5]

However, while the research cited by De Neys and Bonnefon (2013) does indeed indicate that both biased and unbiased reasoners are *able* to detect the conflict between base-rates and stereotypes, for example, little research has directly compared reasoners based on the *extent* of Type 2 processing increase as a function of conflict detection (but see Pennycook et al., 2014a). Do relatively intuitive individuals (i.e., those who are relatively biased) engage in comparable levels of Type 2 processing in the face of conflict as reflective individuals? While it may be the case that intuitive people are able to efficiently detect conflict during reasoning, as suggested by De Neys and Bonnefon, it may also be the case that this conflict detection does not engender much Type 2 processing relative to more analytic individuals. This is an open question that speaks directly to the extent of cognitive processing differences that arise as a function of individual differences.

[5] The use of the term "bias" here refers to participants who scored lower (relatively "biased") or higher (relatively "unbiased") than the median on the given reasoning task (De Neys, 2012). As such, relative bias level, as used by De Neys and colleagues, does not map directly onto either cognitive ability or style. Indeed, given that both style and ability are typically predictive of performance on the tasks used by De Neys and colleagues, "bias" likely reflects a combination of both, depending on the task.

1.6. Current work

The utility of dual-process theory is tied largely to the ability to predict when Type 2 processing will be engaged. Here, we have developed a three-stage model of reasoning and applied it to an illustrative class of reasoning problems. Although the model is consistent with a relatively large body of extant research, there are a number of components that must be empirically tested. Here we investigate two core claims derived from the model: (1) Conflict detection is a key determinant of analytic engagement, but sometimes fails, and (2) conflict detection and cognitive decoupling (i.e., expending additional effort to override an intuitive response in lieu of an alternative) are dissociable sources of analytic thinking. Secondarily, we investigate whether responsiveness to conflicts is subject to individual differences.

To do this, we develop a paradigm that is suitable for measuring subtle increases in Type 2 processing. This paradigm uses base-rate problems which, as outlined above, are of particular interest because they reliably elicit RT increases presumed to result separately from conflict detection and cognitive decoupling processes. To reiterate, participants spend more time on problems that contain a conflict between base-rate and stereotype relative to non-conflict control problems (De Neys & Glumicic, 2008; Pennycook, Fugelsang, et al., 2012). Importantly, this RT increase is evident for both stereotype (IR_1) *and* base-rate (IR_2) responses. As discussed, these RT increases should reflect different processes in the three-stage model. The RT increase for stereotypical responses relative to non-conflict problems is reflective of successful conflict monitoring because such cases reflect sensitivity to IR_2 even when IR_1 is the chosen response (De Neys et al., 2008). Presumably, the additional time is used to rationalize IR_1. In contrast, following previous research that indicates that stereotypical information is a highly salient source of intuitive responses (see Barbey & Sloman, 2007), the RT increase for base-rate responses relative to non-conflict problems should reflect the use of Type 2 processing to *rethink* or decouple from the initial stereotypical response (IR_1), leading to the base-rate response (IR_2). This process should take additional time because IR_1 must be inhibited or suppressed in lieu of the alternative.

Response time is a crucial measure for the current purposes as our focus is on measuring relative *increases* in Type 2 processing as a function of conflict detection and decoupling. Given the presumption that Type 2 processing is typically slower and more resource demanding than Type 1 processing, longer RTs in an experimental condition are thought to reflect an increased level of Type 2 engagement (e.g., Thompson et al., 2011). However, RTs are also notoriously noisy. This is particularly true for typical base-rate problems as mean RTs typically range from 10 to 25 s (De Neys & Glumicic, 2008; Pennycook, Fugelsang, et al., 2012). Thus, we developed a rapid-response version of the base-rate task wherein participants are presented with the individual components of traditional base-rate problems in succession (Pennycook et al., 2014a). In lieu of the long stereotypical descriptions (see above example), participants are presented with a single trait (e.g., "kind") that is strongly diagnostic of one group (e.g., nannies) but not the other (e.g., politicians). This allowed us to decrease extraneous variance due to reading times, increase reliability by including a relatively large number of items, and easily manipulate components of the items across conditions and experiments.

2. Experiment 1

The goal of Experiment 1 was twofold. First, it is necessary to establish the rapid-response paradigm by replicating two key effects.[6] Specifically, participants should take longer for incongruent relative to congruent problems for both stereotypical responses (reflecting successful conflict detection) and base-rate responses (reflecting successful cognitive decoupling).

In contrast to previous models, our three-stage model highlights the possibility that conflict detection may sometimes fail even if two Type 1 outputs are successfully cued by the stimulus. This is consistent with a recent experiment by Mevel et al. (2015), wherein 44% of the sample did not have

[6] Although the rapid-response paradigm has been used in a previous study (Pennycook et al., 2014a), the focus of that paper was on religiosity and the effects with respect to the base-rate task were not reported in detail.

decreased confidence for biased responses to conflict relative to non-conflict ratio bias problems. Thus, although the majority of the sample displayed some evidence that they recognized the inherent conflict in the ratio bias problems (thereby decreasing their confidence in their judgments), a sizable proportion of the sample may have failed to detect the conflict altogether. Indeed, those participants who apparently failed to detect the conflict had lower accuracy on the task and, among the participants who demonstrated a detection effect, there was a positive correlation between the size of the effect and accuracy. These findings suggest that categorical errors in conflict detection are not uncommon and play a role in biased responding. However, accuracy was quite high in their experiment, reaching well over 70% for the participants who ostensibly failed to detect the conflict. Biased responding is typically far more common for base-rate problems and, as a consequence, the rapid-response base-rate task should serve as a strong further test of potential conflict detection failure. Specifically, participants who are highly biased (as indexed by a large proportion of stereotypical responses) should be less likely to differentiate between congruent and incongruent base-rate problems, leading to an absence of a conflict detection effect (i.e., no RT increase for stereotypical responses to incongruent problems relative to congruent). If cases where there is no conflict detection effect are isolated primarily among participants who are highly biased, as opposed to randomly distributed throughout the sample, then we will be justified in calling these conflict detection *failures* and not just random noise in the sample.

2.1. Method

2.1.1. Participants

Sixty-two University of Waterloo undergraduates volunteered to take part in the study in return for partial course credit (16 male, 46 female, M_{age} = 20.95, SD_{age} = 5.46). A subset of this data was reported previously by Pennycook et al. (Experiment 3; 2014a). These data were not analyzed until the full sample was completed. All dependent variables relevant to our target research questions that were analyzed for this experiment are reported below. Participants also completed a religiosity measure in a separate session as a part of a mass-testing survey (see Pennycook et al., 2014a). All manipulations are reported in the method section.

2.1.2. Materials

One-hundred thirty-two items were created using a large online pretest. For this, 86 University of Waterloo undergraduates (28 male, 58 female, M_{age} = 20.6, SD_{age} = 4.06) were given a list of 50 groups of people and asked to select 2 out of 30 personality traits (which was taken from a larger group of 60 total traits, counterbalanced across 2 conditions) that most strongly reflected the prototypical member of the group. The groups primarily consisted of different salient professions (e.g., clown, doctor, etc.) that were likely to be associated with stereotypes. We also included select non-profession groups (i.e., men, women, girls, boys, rich people, poor people, 16 year olds, 50 year olds). The personality traits were well-known stereotypes (e.g., dishonest, punctual, tidy, etc.; see Novemsky & Kronzon, 1999 for a similar strategy). At the end of the pretest, participants were asked two follow-up questions and reminded that they will receive their credit regardless of how they respond. These two questions were: (1) Did you follow the instructions for the above survey, and (2) did you answer the above questions randomly. In total, 7 participants who answered negatively for the first question and/or affirmatively for the second were excluded from further analysis.

To determine which groups were associated with opposing stereotypes, we transposed the data such that the rankings for each of the 60 personality traits were listed for each of the 50 groups. We then investigated the correlations among groups and isolated the top 8 negative correlations for each group (for example, engineer shared opposing stereotypes with groups such as hippy, girl, and clown). We then created 66 sets of groups with opposing stereotypes with the goal of limiting repetition of groups across sets (i.e., each group was paired with more than one other group, but never more than three). Accompanying personality traits were selected based on relative ranking for groups in each set. So, for example, "kind" was selected by 18 participants for nanny and by 1 participant for

politician whereas "dishonest" was selected by 20 participants for politician and by 0 participants for nanny. Finally, we created 2 items for each set using the 2 personality traits, resulting in 132 items in total. However, each set was only presented once per block. For example, nanny and politician were paired with "kind" in the first block and "dishonest" in the second block. This allowed us to counterbalance congruency across blocks.

As in previous research (e.g., De Neys & Glumicic, 2008), we used three extreme base-rate probabilities an equal number of times across trials: 995/5, 996/4, 997/3. Participants received 66 congruent and 66 incongruent problems. Dependent measures included response time and the proportion of base-rate responses. For incongruent problems, response times for stereotype and base-rate responses were included as separate measures in order to index conflict detection and cognitive decoupling respectively.

2.1.3. Procedure

At the beginning of the experiment, participants read the following instructions:

> "In a big research project a large 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 a personality trait for this randomly chosen participant. You'll also get information about the composition of the population groups tested in the study in question.
> You'll be asked to indicate to which population group the participant most likely belongs."
> "Please answer the problems as quickly and accurately as possible. Once you've made up your mind you must enter your answer ('a' or 'b') [corresponding to 'z' and 'm' keys using stickers] immediately and then the next problem will be presented.
> Please feel free to ask any questions that you have."

Participants were then given specific information for each step of the procedure for a practice item. After completing 2 practice items, participants went through two blocks of 66 items each. The procedure for a single item can be seen in Fig. 2.

2.2. Results

2.2.1. Choice proportion for high base-rate alternative

One participant who only chose the response consistent with both base-rate and stereotype 42% of the time on congruent items was removed from analysis; all other participants scored 80% or higher and 90.2% of the sample scored 90% or higher on congruent items. For incongruent problems, the proportion of base-rate responses is the inverse of the proportion of stereotype responses (i.e., .44 base-rate responses = .56 stereotype responses). A large decrease in the proportion of base-rate responses was evident for incongruent relative to congruent items (see Table 1), $t(60) = 11.66$, $SE = .04$, $p < .001$, $d = 1.49$.

2.2.2. Response time

We analyzed both raw response times (RTs) and RTs following a conversion to $\log^{10}$. Outlying raw RT's (3+ *SD*) were excluded prior to our calculation of RT (but not log RT) cell means, representing 1.9% of the data. RTs for congruent items that were inconsistent with both base-rate and stereotype were excluded from analysis (see De Neys & Glumicic, 2008; Pennycook, Fugelsang, et al., 2012 for a similar analytic procedure). As a result, RTs were entered into a repeated measures ANOVA with 3 levels (responses consistent with base-rate/stereotype for congruent items, responses consistent with base-rate for incongruent items, responses consistent with stereotype for incongruent items). A total of 4 participants were not entered into the ANOVA as they gave only stereotypical responses ($N = 3$) or only base-rate responses ($N = 1$) for incongruent items and therefore did not contribute data to each cell of the design. Descriptive statistics for all dependent variables can be found in Supplementary materials (Table S1).

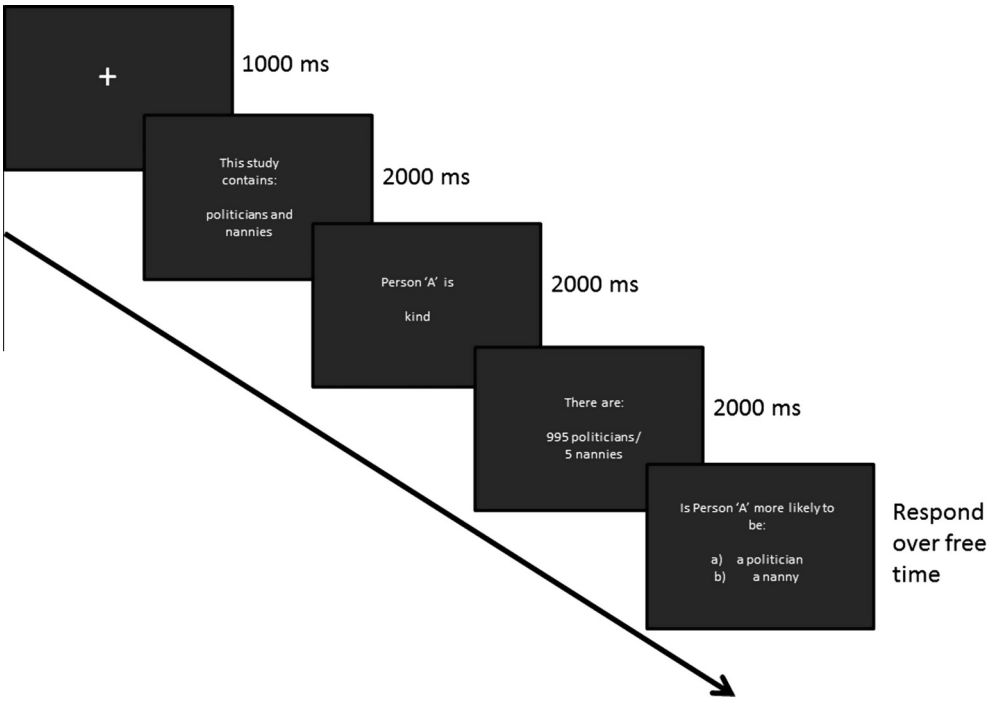

Fig. 2. Procedure for rapid-response base-rate task.

Table 1
Mean choice proportion for high base-rate alternative as a function of problem type for Experiments 1 and 2.

		Choice proportion	
		Congruent	Incongruent
Experiment 1	Extreme base-rates	.97 (.05)	.44 (.35)
Experiment 2	Moderate base-rates	.96 (.04)	.28 (.30)
	Extreme base-rates	.98 (.03)	.49 (.38)

Note: Standard deviations are listed in brackets.

There was a main effect of congruency on RT, $F(1.3, 74.3) = 21.36$, $MSE = 222754.7$, $p < .001$, $\eta^2 = .28$ (see Table 2) and log RT, $F(1.2, 70.0) = 31.09$, $MSE = .04$, $p < .001$, $\eta^2 = .35$.[7] Planned follow-up t-tests revealed significant differences between congruent and incongruent base-rate responses, $t(59) = 7.65$, $SE = 70.41$, $p < .001$, $d = .99$, and congruent and incongruent stereotype responses, $t(57) = 5.72$, $SE = 68.27$, $p < .001$, $d = .75$. Incongruent base-rate responses and incongruent stereotype responses did not significantly differ, $t(56) = 1.41$, $SE = 115.66$, $p = .163$, $d = .19$, though this analysis was marginal with log RT's, $t(57) = 1.99$, $SE = .05$, $p = .063$, $d = .25$. These results indicate that the RT increase for incongruent relative to congruent items was evident for both base-rate and stereotypical responses, but that it tended to be somewhat larger for base-rate responses. These results closely match those found using traditional base-rate problems (De Neys & Glumicic, 2008; Pennycook, Fugelsang, et al., 2012).

[7] Values were adjusted using the Greenhouse–Geisser correction in this and the following experiments when the sphericity assumption was violated.

Table 2
Mean response time (in milliseconds) as a function of problem type and response (either consistent with stereotype or base-rate) for Experiments 1 and 2. Difference (in milliseconds) between incongruent (stereotypical or base-rate response) and congruent response times are in brackets.

		Congruent	Incongruent	
			Stereotypical (diff)	Base-rate (diff)
Experiment 1	Extreme base-rates	696 (38)	1095* (79)	1258* (84)
Experiment 2	Moderate base-rates	778 (60)	898* (137)	1470* (193)
	Extreme base-rates	787 (59)	1385* (135)	1504* (190)

* Significance at a .05 level for the incongruent–congruent response time comparison. Standard error is listed in brackets.

2.2.3. Individual differences

Next we turn to the prediction that individual differences would reveal categorical failures in conflict detection. As an initial step, we correlated the proportion of base-rate responses for congruent and incongruent problems with several RT measures (see Table 3). The goal here is to investigate if there is an overall correlation between biased responding and RT for incongruent problems as a means to justify our isolation of potential *categorical* conflict detection failures. If there are cases of conflict detection failure, we expect them to be associated with high levels of biased responding, which implies a positive correlation between RT for stereotypical responses and the proportion of base-rate responses (i.e., the inverse of biased responding). An alternative possibility is that cases where there is no difference between RT for stereotypical responses and the congruent baseline simply represent random noise in the sample and not genuine failures of conflict detection. Anything but a clear association between presumed detection failures and high levels of bias would support this possibility.

Given the skew for raw RTs, we use the log RT difference scores for this analysis. As above, the participants who did not contribute data to every cell of the design (i.e., those who gave all base-rate or stereotypical responses) were excluded from analysis. Moreover, we included RT for base-rate responses in this analysis for completeness. The results can be found in Table 3.

First, as expected, the proportion of base-rate responses for congruent problems did not correlate with any other measure. Moreover, RT for congruent problems was not associated with the proportion of base-rate responses for incongruent problems, though there was a significant correlation between RT for congruent and incongruent problems. These results indicate the RT for congruent problems is a good baseline, which is consistent with previous research (Pennycook, Fugelsang, et al., 2012; Pennycook et al., 2014a). As a consequence, we subtracted the RT for congruent problems from the RTs for incongruent stereotypical and base-rate responses.[8] The theorized increase in analytic processing due to conflict detection is indexed by the RT increase for incongruent *stereotype* responses relative to the congruent baseline whereas the theorized increase in analytic processing due to cognitive decoupling is indexed by the RT increase for incongruent *base-rate* responses relative to the congruent baseline (De Neys & Glumicic, 2008; Pennycook, Fugelsang, et al., 2012).

The proportion of base-rate responses for incongruent problems was strongly *positively* correlated with RT for stereotypical responses to incongruent problems and strongly *negatively* correlated with RT for base-rates responses to incongruent problems. Moreover, these correlations increased in magnitude when RT for the congruent baseline was subtracted out to create the 'conflict detection' and 'cognitive decoupling' indices. These indices were also strongly negatively correlated. The scatterplots for the correlations between the proportion of base-rate responses to incongruent problems and the conflict detection and cognitive decoupling effects are overlaid in Fig. 3. Each unit of analysis in Fig. 3 represents a participant (i.e., one circle and one triangle for each participant).

Fig. 3 allows for an inspection of the categorical failures of conflict detection. Specifically, there is a cluster of participants who gave a large majority of stereotypical responses and who did not spend any

[8] To establish the reliability of these difference scores, we broke the trials up into 4 sets of 33 observations (i.e., the randomized items for each of the 2 blocks were each split in half based on the order of presentation). The conflict detection effect across these 4 sets had good reliability (α = .71). The cognitive decoupling effect had acceptable reliability (α = .65).

Table 3
Pearson product–moment correlations between the proportion of base-rate responses and RTs in Experiment 1. Base-rate % = proportion of base-rate responses. Conflict detection refers to the difference in RT between incongruent stereotypical responses and congruent items. Cognitive decoupling refers to the difference in RT between incongruent base-rate responses and congruent items. Coefficients in bold are significant, $p < .05$. $N = 58$.

	1	2	3	4	5	6	7
1. Incongruent – base-rate prop.	–	.16	.04	**.65**	**.84**	**−.52**	**−.70**
2. Congruent – base-rate prop.		–	−.09	.09	.21	−.06	<−.01
3. Congruent RT			–	**.67**	−.10	**.61**	−.02
4. Incongruent stereotype RT				–	**.68**	.14	**−.37**
5. Conflict detection (RT diff)					–	**−.43**	**−.47**
6. Incongruent base-rate RT						–	.78
7. Cognitive decoupling (RT diff)							–

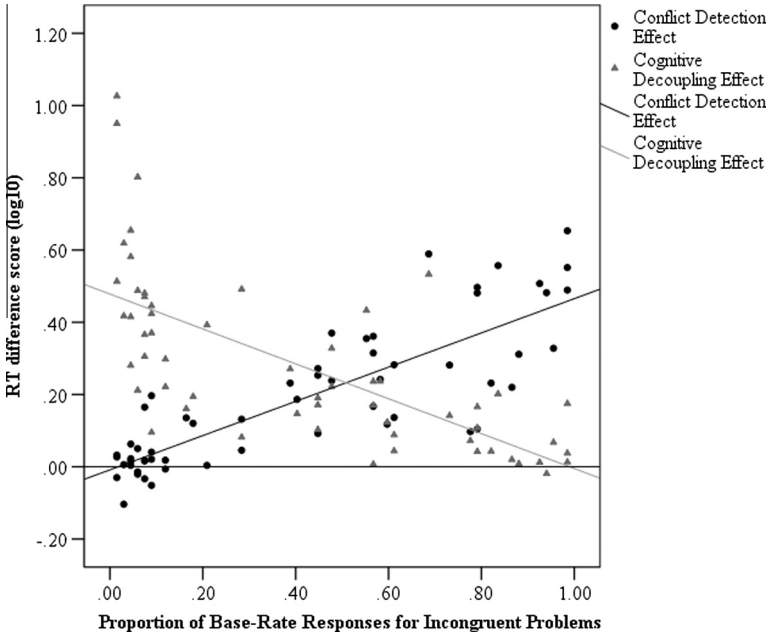

Fig. 3. Scatterplot of mean response time differences ($\log^{10}$) and the proportion of base-rate responses for incongruent problems in Experiment 1. Conflict detection refers to the difference in RT between incongruent stereotypical responses and congruent items. Cognitive decoupling refers to the difference in RT between incongruent base-rate responses and congruent items. Each unit of analysis represents a participant (i.e., one circle and one triangle for each participant). Lines show regression of RT difference scores ($\log^{10}$) on proportion of base-rate responses.

additional time (relative to the congruent baseline) doing so. A *post hoc* investigation of the conflict detection difference scores revealed that 10 out of 58 participants (17.2%) had a negative raw RT difference between congruent and incongruent stereotypical responses (for a similar approach, see Mevel et al., 2015).[9] Importantly, these participants were particularly biased in that they only responded according to the base-rate 5.8% of the time. This was more than 40% lower than the remainder of the sample ($M = 48.7\%$), $t(56) = 4.14$, $p < .001$. We computed the error rate for congruent problems (i.e., selections inconsistent with both base-rate and stereotypes) for these participants ($M = 5.4\%$) and there was

[9] We used raw RT instead of log RT difference scores for this analysis because there is no concern for outliers (given the focus on the small and slightly negative difference scores) and the raw RTs are easier to interpret.

no significant difference between congruent and incongruent problems, $t < 1$. Thus, not only did these participants show no evidence of conflict detection as measured by RT, but their responses revealed no influence of the base-rate information whatsoever. These data indicate that the most biased subset of the sample may have largely failed to detect the conflict between base-rate and stereotype. Indeed, they may have been particularly biased precisely *because* the conflict failed to cue analytic thinking.

Despite these findings, it should be noted that 82.8% of the sample had a positive RT difference between congruent and incongruent problems for stereotypical answers (ranging from 5 to 2323 ms). Note, however, that this is a liberal estimate of the proportion of the sample who successfully detected the conflict between base-rates and stereotypes as it includes individuals with very small overall effects. For example, 14 participants (24.1%) had a RT difference that was 100 ms or less, and these participants also had a low proportion of base-rate responses ($M = 13.1\%$) relative to the overall mean of 44% (see Table 1). Nonetheless, even if this group is assumed to not have successfully detected the conflict, the majority of the sample still remains (58.7%; incidentally, a number that closely matches Mevel et al., 2015). Thus, despite the variation in the conflict detection effect, there is evidence that the majority of the sample *did* increase Type 2 processing following successful conflict detection.

2.3. Discussion

The results of Experiment 1 indicate that individual differences play a substantial role in the degree of Type 2 engagement following the presentation of conflicting information in a reasoning paradigm. Biased responding was associated with a smaller increase in RT for stereotypical responses to incongruent relative to congruent problems, potentially indicating that less biased individuals are more responsive to conflict (see also Pennycook et al., 2014a). Moreover, a non-trivial proportion of the sample (17.2%) actually took longer for congruent than incongruent problems and gave base-rate responses to incongruent problems at roughly the same rate (5.8%) as they gave the patently incorrect response to congruent problems (5.4%). This concordance between biased responding and categorical failures of conflict detection suggests that conflict detection is *not* perfectly efficient. However, this pattern of results was isolated to a minority of participants and, as a consequence, conflict detection failures of this type appeared to be relatively rare for this task.

Somewhat surprisingly, less biased individuals actually took *less* time to respond than more biased individuals in cases where the stereotypical response was successfully overridden. To our knowledge, these results represent the first evidence that this association is not simply defined by an increase in RT among less biased individuals (see De Neys & Bonnefon, 2013). This may indicate that a lower level of biased responding is associated with efficiency in cognitive decoupling; in other words, that failures in decoupling ("inhibition failures"; see De Neys, 2012) are an important source of biased responding. There was no evidence that base-rate responses are simply the result of a "guess" that results from having two equally intuitive choices. This finding is consistent with a recent set of experiments using syllogisms (e.g., "No fruits are sour, All lemons are fruits, Therefore, no lemons are sour."). In these experiments Svedholm-Häkkinen (2015) replicated and extended the key finding from De Neys and Franssens (2009): i.e., that successfully inhibiting a belief-based intuitive response (i.e., decoupling) for conflict problems (using both syllogisms and base-rate problems) leads to impaired memory for the relevant belief on a later lexical decision task. Svedholm-Häkkinen found that particularly skilled participants (i.e., those high in cognitive ability and/or who have an analytic thinking disposition) did *not* show evidence of belief inhibition following a syllogistic reasoning task. This supports the idea that some reasoners may be particularly efficient at cognitive decoupling.

Future research could isolate the underlying factors that determine decoupling efficiency in base-rate tasks such as the one employed here. Since cognitive decoupling requires both the inhibition of the initial response and a search for alternatives (Stanovich, 2009a; Stanovich & West, 2008), it is often strongly related to individual differences in fluid intelligence (e.g., Unsworth & Engle, 2005, 2007). Given this work, one possibility is that cognitive ability contributes to the suppression of stereotypical responses in lieu of base-rate responses. However, it is important to keep in mind that the increases in analytic processing that are being probed in the rapid-response base-rate paradigm are small. The increase in time spent for base-rate responses to incongruent relative to congruent

problems was 562 ms (see Table 2). This is a large increase in relative terms (participants take almost twice as long for incongruent relative to congruent problems), but the task of decoupling from the intuitive stereotype is only so complex as to require around half a second to complete. Also, it is possible that the efficiency of cognitive decoupling in this task relates to higher levels of statistical knowledge (i.e., having sufficient mindware to easily override the stereotypical response) or less intuitive stereotypical responses (i.e., intuitive responses cued by stereotypes are less compelling and therefore easier to override). Though the foregoing does not bear directly on the proposed three-stage model as a general account of analytic engagement, it is nonetheless an interesting area for future research on base-rate neglect.

3. Experiment 2

In Experiment 1, we used the rapid-response base-rate task to replicate a set of key results. Moreover, we found evidence that particularly biased participants may be biased, in part, because they failed to detect the conflict between base-rates and stereotypes. Finally, further investigation of the correlation between the proportion of base-rate responses and the conflict detection and cognitive decoupling effects seemed to reveal a dissociation between these two potential sources of analytic thinking. That is, biased responding was associated with small conflict detection effects and large cognitive decoupling effects. However, this result should be interpreted with caution as it may be the case that some people enter the experiment with a stronger or weaker bias toward stereotypes or base-rates and, as a consequence, they simply respond faster with their dominant response.[10] Thus, in the next experiment, we look to test this dissociation experimentally.

As discussed above, Pennycook, Fugelsang, et al. (2012) found that the probability of conflict detection (as indexed by RT increases for stereotypical responses) for traditional base-rate problems is substantially diminished when base-rates are moderate (e.g., 70/30) relative to when they are extreme (e.g., 995/5). Indeed, there was no evidence of conflict detection whatsoever in any of the three experiments that did not include extreme base-rates. Although the comparison was across experiments, this could be thought of as the first reported *manipulation* of conflict detection in a reasoning paradigm. Crucially, there was a significant RT increase for base-rate responses to incongruent problems relative to congruent in all five experiments, indicating that cognitive decoupling was not affected by base-rate extremity. There was no cross-experiment test comparing the extent of the increase, however, so this can only be considered preliminary evidence that conflict detection and cognitive decoupling are separable sources of increases in Type 2 processing. As such, we introduced base-rate extremity as a manipulation in Experiment 2.

As in Experiment 1, we will also investigate the correlation between proportion of base-rate responses and the RT increases that we have attributed to conflict detection and cognitive decoupling processes. Bias susceptibility should be positively correlated with RT for stereotypical responses (conflict detection) and negatively correlated with RT for base-rate responses (decoupling) for both moderate and extreme base-rates. However, the proportion of *categorical* failures of conflict detection should be higher for moderate than extreme base-rates.

3.1. Method

3.1.1. Participants

Sixty University of Waterloo undergraduates volunteered to take part in the study in return for partial course credit. Participants were randomly assigned to a moderate (6 male, 24 female, M_{age} = 19.2, SD_{age} = 1.4) or an extreme (6 male, 24 female, M_{age} = 20.1, SD_{age} = 2.6) base-rate condition. These data were not analyzed until the full sample was completed. All dependent variables that were analyzed for this experiment are reported below and all manipulations are reported in the method section (see Footnote [6]).

[10] We would like to thank Jonathan Evans for alerting us to this possibility.

3.1.2. Materials and procedure

The materials and procedure were identical to Experiment 1 with the exception that half of the participants were given base-rate problems with moderate base-rates. We used three moderate base-rate probabilities an equal number of times across trials: 700/300, 710/290, 720/280.[11]

3.2. Results

3.2.1. Choice proportion for high base-rate alternative

All participants chose the response consistent with both base-rate and stereotype for congruent problems more than 80% of the time. We entered the proportion of base-rate responses into a 2 (Congruency: incongruent, congruent) × 2 (Extremity: moderate, extreme) mixed ANOVA (see Table 1). There was a large overall decrease in proportion of base-rate responses for incongruent relative to congruent items, $F(1,58) = 179.28$, $MSE = .057$, $p < .001$, $\eta^2 = .76$. There was also a between subject difference wherein the proportion of base-rate responses was lower for the moderate base-rate condition ($M = .62$) relative to the extreme base-rate condition ($M = .74$), $F(1,58) = 6.12$, $MSE = .062$, $p = .016$, $\eta^2 = .10$, and an interaction between congruency and condition, $F(1,58) = 4.89$, $MSE = .057$, $p = .031$, $\eta^2 = .08$, indicating that the difference between conditions was larger for incongruent relative to congruent items (see Table 1).

3.2.2. Response time

As above, we analyzed both raw response times (RTs) and RTs following a conversion to $\log^{10}$. Outlying raw RT's (3+ *SD*) were excluded prior to our calculation of cell means, representing 0.6% of the data. We entered the resulting RTs into a 3 (Congruency: incongruent base-rate, incongruent stereotype, congruent) × 2 (Condition: moderate, extreme) mixed ANOVA. A total of 5 participants were not entered into the ANOVA as they gave only stereotypical responses ($N = 4$) or only base-rate responses ($N = 1$) for incongruent items and therefore did not contribute data to each cell of the design. Mean RTs can be found in Table 2. Descriptive statistics for all dependent variables can be found in Supplementary materials (Tables S2 and S3).

There was a main effect of congruency on RT, $F(1.3,68.9) = 17.11$, $MSE = 612984.5$, $p < .001$, $\eta^2 = .24$ (see Table 2) and log RT, $F(1.2,65.9) = 34.71$, $MSE = .06$, $p < .001$, $\eta^2 = .39$. There was no main effect of condition for either RT, $F(1,53) = 1.60$, $MSE = 810871.4$, $p = .212$, $\eta^2 = .03$, or log RT, $F(1,54) = .25$, $MSE = .13$, $p = .622$, $\eta^2 < .01$. However, there was a marginal interaction for RT, $F(2,106) = 2.49$, $MSE = 398458.1$, $p = .088$, $\eta^2 = .05$, and significant interaction for log RT, $F(2,108) = 5.81$, $MSE = .033$, $p = .004$, $\eta^2 = .10$. To further investigate this interaction, we computed the two RT difference scores in the same manner as in Experiment 1: (1) the difference between RTs for incongruent stereotypical and congruent, and (2) the difference between RTs for incongruent base-rate and congruent.[12]

As is evident from Table 2, the difference between the size of the RT difference between stereotypical responses to incongruent problems and congruent (control) problems was, as predicted, larger when participants were presented with extreme base-rates relative to moderate ones, RT: $t(54) = 3.29$, $SE = 146.5$, $p = .002$, $d = .88$, log RT: $t(54) = 3.25$, $SE = .04$, $p = .002$, $d = .87$. There was no between subject difference for the cognitive decoupling effect, RT: $t(57) = .13$, $SE = 225.1$, $p = .895$, log RT: $t(58) = 1.32$, $SE = .06$, $p = .191$. This finding replicates the pattern of results found by Pennycook, Fugelsang, et al. (2012). However, in contrast to that experiment, all RT difference scores in the current experiment were greater than zero, all t's > 2.8, all p's < .01. This indicates that

[11] Pennycook, Fugelsang, et al. (2012) used moderate base-rates with similar probabilities (i.e., ~70%), but on a different scale (i.e., 70/30 instead of 700/300). Theoretically, this should not make a difference. To test this assumption we also included two additional moderate base-rate conditions ($N = 60$, 12 male, 48 female, $M_{age} = 20.0$, $SD_{age} = 2.0$): 7/3, 8/2 and 70/30, 71/29, 72/28. As expected, these conditions did not differ from the reported moderate base-rate condition (i.e., 700/300, etc.) for any RT measure, all F's < 1. Nor were there any differences in proportion of base-rate responses: incongruent, $F < 1$; congruent, $F(2,87) = 2.16$, $p = .122$. We excluded these conditions from the primary analysis to facilitate exposition.

[12] We ran the same reliability analysis as in Experiment 1 (see Footnote [8]). The conflict detection effect had acceptable reliability in the moderate base-rate condition ($\alpha = .63$) and good reliability in the extreme base-rate condition ($\alpha = .76$). The cognitive decoupling effect had good reliability in the moderate base-rate condition ($\alpha = .78$) and acceptable reliability in the extreme base-rate condition ($\alpha = .67$).

participants in the moderate base-rate condition were, on the aggregate, able to successfully detect the conflict in the rapid-response version of the task, albeit, as noted above, the overall responsiveness to conflict was lower for moderate than with extreme base-rates. This may reflect greater sensitivity for the rapid response task or it may be a result of the much larger number of items in the new task relative to previous versions.

3.2.3. Individual differences

We correlated the proportion of base-rate responses with our RT measures separately for the two conditions. As in Experiment 1, we correlated the RTs after $\log^{10}$ conversion and excluded participants who gave all base-rate or all stereotypical responses. As is evident from Table 4, the results from Experiment 1 were replicated in the extreme base-rate condition (below-diagonal of Table 4) and extended in the moderate base-rate condition (above-diagonal of Table 4). The conflict detection effect was strongly *positively* correlated with the proportion of base-rate responses for both moderate ($r = .82$) and extreme ($r = .77$) base-rates, and the cognitive decoupling effect was strongly *negatively* correlated with the proportion of base-rate responses for both moderate ($r = -.55$) and extreme ($r = -.84$) base-rates. Moreover, inspection of the scatterplots (see Fig. 4) reveals that the decreased

Table 4
Pearson product–moment correlations between the proportion of base-rate responses and RTs in Experiment 2. Correlations for the moderate base-rate condition are displayed on the above-diagonal. Correlations for extreme base-rate condition are displayed on the below-diagonal. Base-rate prop = proportion of base-rate responses. Conflict detection refers to the difference in RT between incongruent stereotypical responses and congruent items. Cognitive decoupling refers to the difference in RT between incongruent base-rate responses and congruent items. Coefficients in bold are significant, $p < .05$. $N = 28$ (in each condition).

	1	2	3	4	5	6	7
1. Incongruent – base-rate prop.	–	−.03	−.17	.29	**.82**	**−.56**	**−.55**
2. Congruent – base-rate prop.	.18	–	−.03	.09	.04	.11	.16
3. Congruent RT	.11	−.10	–	**.84**	−.21	**.61**	.16
4. Incongruent stereotype RT	**.59**	<.01	**.74**	–	.34	.28	−.16
5. Conflict detection (RT diff)	**.77**	.13	.10	**.75**	–	**−.41**	−.36
6. Incongruent base-rate RT	**−.47**	−.19	**.74**	.29	−.33	–	.88
7. Cognitive decoupling (RT diff)	**−.84**	−.17	.04	−.37	**−.63**	**.71**	–

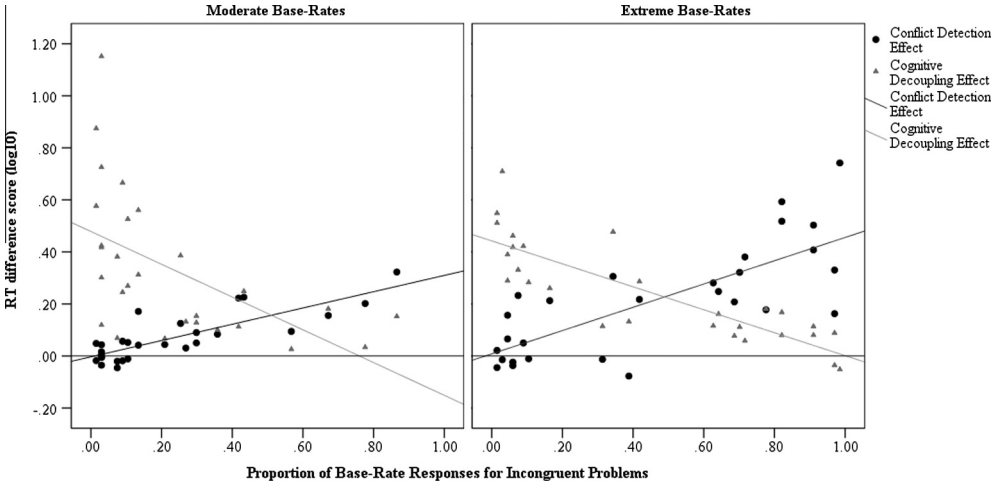

Fig. 4. Correlations between mean response time differences ($\log^{10}$) and the proportion of base-rate responses for incongruent problems in Experiment 2. Conflict detection refers to the difference in RT between incongruent stereotypical responses and congruent items. Cognitive decoupling refers to the difference in RT between incongruent base-rate responses and congruent items. Each unit of analysis represents a participant (i.e., one circle and one triangle for each participant). Lines show regression of RT difference scores ($\log^{10}$) on proportion of base-rate responses.

proportion of base-rate responses in the moderate condition (Moderate = .28; Extreme = .49; see Table 1) may be partially accounted for by an increased number of participants who failed to detect the conflict between base-rates and stereotypes. The conflict detection difference scores (using raw RT) revealed that, in the moderate base-rate condition, 8 out of 28 participants (28.6%) had a negative RT difference between congruent and incongruent stereotypical responses (Mean proportion of base-rate responses = 6.2%). A further 12 participants (42.9%) in the moderate condition had a positive conflict detection effect that was smaller than 100 ms (Mean proportion of base-rate responses = 17%), indicating that the majority of the sample (71.5%) either had a negative or very slightly positive conflict detection effect. In the extreme base-rate condition, 5 out of 28 participants (17.9%) had a negative RT difference (Mean proportion of base-rate responses = 11%) and 5 (17.9%) had a positive difference that was less than 100 ms (Mean proportion of base-rate responses = 11.3%). Thus, in contrast to the moderate base-rate condition and consistent with Experiment 1, the majority of the sample (64.2%) had a positive conflict detection effect that was reasonably robust. This difference was significant using a Chi-Square analysis, $\chi^2(1, N = 56) = 7.18$, $p = .007$. Although 100 ms is an arbitrary cut off, these results nonetheless suggest that the overall decrease in the conflict detection effect for moderate relative to extreme base-rates may at least be partially the result of categorical conflict detection failures.

3.3. Discussion

Participants took longer for stereotypical responses to incongruent relative to congruent problems in both the extreme and moderate base-rate conditions; however, the extent of this difference was substantially smaller in the latter case (i.e., 120 ms for moderate compared to 598 ms for extreme; see Table 2). In contrast, base-rate extremity had no effect on the RT increase for base-rate responses to incongruent relative to congruent problems (i.e., 692 ms for moderate compared to 717 ms for extreme; see Table 2). Thus, as hypothesized, extreme base-rate probabilities appeared to have increased the responsiveness to cognitive conflict without affecting cognitive decoupling. This increased the amount of time spent rationalizing the initial stereotypical response (IR_1) following successful conflict detection, but had no effect on the amount of time taken to override it in lieu of the base-rate response (IR_2). This serves as an initial justification of our separation of these components in the three-stage model.

There was also evidence for increased conflict detection failures among relatively more biased participants, particularly when given moderate base-rates. It appears that failures of conflict detection are, at the very least, not uncommon. Further, replicating Experiment 1, the conflict detection effect was larger for less biased relative to more biased individuals; a finding that held for both moderate and extreme base-rates. We should reiterate, however, that these results should be interpreted with caution as the current experiment cannot rule out the possibility that participants enter the experiment with a differential predisposition toward stereotype or base-rate information. Note, however, that this explanation cannot account for the effect of base-rate extremity on RT increases for stereotypical responses (conflict detection), nor the lack of this effect for base-rate responses (decoupling). This finding is somewhat counterintuitive as manipulating the *base-rates* selectively affected RT for *stereotypical* responses; a finding difficult to interpret without appealing to the possibility that extreme base-rates facilitate conflict detection specifically.

4. Experiment 3

The goal of Experiment 3 was to extend Experiments 1 and 2 in two important ways. The first relates to our use, as per De Neys and colleagues (e.g., De Neys, Moyens, & Vansteenwegen, 2010), of the proportion of base-rate responses to index bias susceptibility. This analysis strategy has the benefit of allowing us to map our results directly on to previous predictions made about conflict detection (e.g., De Neys & Bonnefon, 2013). However, as discussed above, individual differences in cognitive style, in particular, have specific relevance for our three-stage model because cognitive style can be considered an independent top-down source of Type 2 engagement. The first goal of Experiment 3,

therefore, was to investigate whether there is a specific association between the willingness to engage Type 2 processing and Type 2 processing following conflict detection by including a self-report measure of thinking disposition.

There is some preliminary evidence for such an association. Pennycook et al. (2014a) found a positive correlation between the degree of RT increase for stereotypical responses to incongruent relative to congruent problems and both self-report (i.e., the Actively Open-minded Thinking questionnaire; see their Experiment 1) and performance-based (i.e., the Cognitive Reflection Test; see their Experiment 2) measures of cognitive style. However, the authors used traditional base-rate problems with extreme base-rates. Here we employ the more reliable rapid-response task and include moderate base-rates. If cognitive style is associated with increased responsiveness to conflict *in particular*, there should be an association between cognitive style and the RT increase for stereotypical responses (the "conflict detection effect") but not the RT increase for base-rate responses (the "cognitive decoupling effect"). Moreover, the correlation between RT for stereotypical responses should be stronger for moderate relative to extreme base-rates. The underlying assumption here is that cases where cognitive decoupling requires relatively more Type 2 processing – as, for example, among more biased relative to less biased participants – are explained by greater difficulty suppressing or inhibiting the intuitive stereotypical response. Thus, it is a matter of capacity (e.g., processing speed, statistical knowledge), not disposition.

The second goal of Experiment 3 was to manipulate the probability of conflict detection without altering the content of the items. Kahneman and Tversky's (1973) single lawyer–engineer problem was the basis of the base-rate neglect problems developed by De Neys and Glumicic (2008) to investigate conflict detection during reasoning. Although presenting participants with multiple versions of the same problem is necessary for this type of research, responsiveness to the conflict between base-rate and stereotype may increase as the number of problems increases (Kahneman, 2000; Kahneman & Frederick, 2002). Alternatively, it is possible that a large number of items may lead to habituation, thereby diminishing our effects.[13] Thus, we included base-rate extremity as a *within-subject* variable and varied the order of presentation between-subjects. This is a simple manipulation that should have powerful effects. Specifically, under the hypothesis that the responsiveness to cognitive conflict is facilitated by earlier detections within a problem set, RT for stereotypical responses to incongruent problems with moderate base-rates should increase if extreme base-rates are presented in an earlier block of trials. Indeed, presenting extreme base-rates prior to moderate base-rates should greatly diminish the base-rate extremity effect. In contrast, presenting moderate base-rates prior to extreme base-rates should have the opposite effect. That is, the switch from moderate to extreme should make the base-rates highly salient and greatly increase the responsiveness to conflict. As in Experiment 2, this manipulation should have no effect on the amount of time spent decoupling from the stereotypical response (i.e., RT for base-rate responses to incongruent problems will not differ between conditions).

4.1. Method

4.1.1. Participants

Seventy-four University of Waterloo undergraduates volunteered to take part in the study in return for partial course credit. Participants were randomly assigned to either the extreme first, moderate second condition (11 male, 26 female, M_{age} = 20.5, SD_{age} = 1.7) or the moderate first, extreme second condition (13 male, 24 female, M_{age} = 20.5, SD_{age} = 1.6). These data were not analyzed until the full sample was completed. All manipulations are reported in the method section.

4.1.2. Materials and procedure

The materials and procedure for the rapid-response task were identical to Experiment 2 with the exception that base-rate extremity was manipulated within subject. Thus, in total, participants were given 264 items across two blocks, counterbalanced across condition.

[13] We would like to thank Wim De Neys for alerting us to this possibility.

Participants were given a thinking disposition questionnaire consisting of 18 items from the Need for Cognition scale (NFC: Cacioppo, Petty, Feinstein, & Jarvis, 1996) and 41 items from the Actively Open-minded Thinking scales (AOT: Stanovich & West, 2007), presented in a randomly intermixed order. The scales included questions such as "I usually end up deliberating about issues even when they do not affect me personally" (NFC) and "changing your mind is a sign of weakness" (AOT, reverse scored). The overall thinking disposition score was obtained by summing the responses across all items. Each item was scored such that higher scores represented a greater tendency toward analytic thinking. The full scale had good internal consistency: Cronbach's $\alpha = .91$.[14]

4.2. Results

4.2.1. Choice proportion for high base-rate alternative

Five participants who chose the response consistent with both base-rate and stereotype for congruent problems less than 80% of the time were excluded from further analysis (4 when base-rates were moderate and 1 when base-rates were extreme). We entered the proportion of base-rate responses into a 2 (Congruency: incongruent, congruent) × 2 (Extremity: moderate, extreme) × 2 (Condition: moderate first, extreme first) mixed ANOVA (Table 5). There was a decrease in the proportion of base-rate responses for incongruent relative to congruent items, $F(1,67) = 148.54$, $MSE = .11$, $p < .001$, $\eta^2 = .69$. There was also a within subject difference wherein the proportion of base-rate responses was lower for moderate base-rates relative to extreme base-rates, $F(1,67) = 44.49$, $MSE = .014$, $p < .001$, $\eta^2 = .40$, and an interaction between congruency and extremity, $F(1,67) = 42.90$, $MSE = .012$, $p < .001$, $\eta^2 = .39$, indicating that the difference between moderate and extreme base-rates was larger for incongruent relative to congruent items.

There was no between subject difference in overall proportion of base-rate choices and no congruency by order condition interaction, both Fs < 1. However, order condition did interact with extremity, $F(1,67) = 15.43$, $MSE = .014$, $p < .001$, $\eta^2 = .19$, and there was a three-way interaction between congruency, extremity, and condition, $F(1,67) = 17.13$, $MSE = .012$, $p < .001$, $\eta^2 = .20$. As is evident from Table 5, the source of this interaction appears to be a 30% increase in base-rate responses for extreme relative to moderate items in the moderate first condition. To investigate this possibility, we computed a difference score between moderate and extreme items for incongruent problems. An independent samples t-test verified that the difference between moderate and extreme base-rate items was substantially larger when moderate base-rate items were presented *before* extreme base-rate items, $t(67) = 4.09$, $SE = .05$, $p < .001$, $d = .99$. This suggests that the extreme base-rates were made particularly salient when preceded by moderate base-rates, leading participants to actually select the base-rate alternative at a nominally higher rate than the stereotypical option.

4.2.2. Response time

We analyzed both raw response times (RTs) and RTs following a conversion to $\log^{10}$. Outlying raw RT's (3+ SD) were excluded prior to our calculation of cell means, representing 0.9% of the data. We analyzed the data using the same RT difference scores that were computed in Experiments 1 and 2 (see Table 6 for mean RTs). Descriptive statistics for all dependent variables can be found in Supplementary materials (Tables S4–S7).

Dependent variables were entered into separate 2 (Extremity: moderate, extreme) × 2 (Condition: moderate first, extreme first) mixed design ANOVAs. Thirteen participants were not included in the ANOVA because they did not give any stereotypical responses (5 for extreme base-rates, 3 for moderate base-rates, and 5 for both). Replicating Experiment 2, the RT difference between stereotypical responses to incongruent problems and congruent problems (i.e., conflict detection) was marginally larger for extreme base-rates relative to moderate ones for RT, $F(1,54) = 3.96$, $MSE = 369163.4$, $p = .052$, $\eta^2 = .07$, and significantly larger for log RT, $F(1,54) = 10.10$, $MSE = .03$, $p = .002$, $\eta^2 = .16$. There was no between subject effect of order condition for either RT or log RT, Fs < 1. Curiously, there

[14] We also analyzed the correlations separately for the NFC and AOT scales. The results were similar for both scales, though, as one would expect, the correlation coefficients were typically smaller than for the full scale.

Table 5
Mean choice proportion for high base-rate alternative as a function of problem type and condition for Experiment 3.

	Moderate first		Extreme first	
	Congruent	Incongruent	Congruent	Incongruent
Moderate	.96 (.04)	.35 (.34)	.96 (.05)	.41 (.37)
Extreme	.97 (.05)	.65 (.37)	.97 (.05)	.48 (.36)

Note: Standard deviations are listed in brackets.

Table 6
Mean response time (in milliseconds) as a function of problem type, response (either consistent with stereotype or base-rate), base-rate extremity, and order for Experiment 3.

Order condition	Base-rate extremity	Congruent	Incongruent	
			Stereotypical	Base-rate
Moderate–extreme	Moderate	840 (64)	1065 (148)	1215 (155)
Moderate–extreme	Extreme	714 (72)	1291 (169)	1068 (167)
Extreme–moderate	Moderate	624 (64)	862 (148)	1010 (155)
Extreme–moderate	Extreme	811 (72)	1167 (169)	1305 (167)

Note: Standard error is listed in brackets.

was no interaction between extremity and order condition for RTs, $F(1,54) = .70$, $MSE = 369163.4$, $p = .406$, $\eta^2 = .01$, but a significant interaction for log RTs, $F(1,54) = 5.00$, $MSE = .03$, $p = .029$, $\eta^2 = .09$. Further inspection of the mean RTs revealed four outliers (3 *SD*s); two for extreme and two for moderate. With the outliers removed, the interaction between extremity and condition was robust for RT, $F(1,51) = 10.29$, $MSE = 108911.6$, $p = .002$, $\eta^2 = .17$. This also led to a more robust main effect for RT, $F(1,51) = 19.36$, $MSE = 108911.6$, $p < .001$, $\eta^2 = .28$. As is clear from Table 7, the RT difference for stereotypical responses was much larger for extreme base-rates than moderate base-rates if moderate base-rates were presented first, RT (outliers removed): $t(26) = 4.28$, $SE = 113.9$, $p < .001$, $d = .82$; log RT (no outliers removed): $t(27) = 2.94$, $SE = .06$, $p = .007$, $d = .56$. When extreme base-rates were presented first, there was no difference between moderate and extreme base-rates, RT (outliers removed): $t(25) = 1.38$, $SE = 55.4$, $p = .180$, $d = .27$; log RT (no outliers removed): $t(27) = 1.21$, $SE = .02$, $p = .236$, $d = .23$. This finding indicates that presenting participants with extreme base-rates prior to moderate ones increases the responsiveness to conflict for moderate base-rates. In contrast, conflict detection sensitivity for moderate base-rates was very modest when they were presented first.[15] Moreover, it appears that switching from moderate to extreme base-rates made the base-rates highly salient, leading to a very robust conflict detection effect.

The RT difference between incongruent base-rate responses and congruent items (i.e., cognitive decoupling) did not differ as a function of base-rate extremity or condition for either RT or log RT, all Fs < 1. Nor was there an interaction between extremity and condition for either RT or log RT, all Fs < 1 (see Table 7).

4.2.3. Individual differences

Correlations between analytic thinking disposition and major variables are presented in Table 8. Given the skew for raw RTs, we use the log RT difference scores for this analysis. The log RT difference score for stereotypical responses was positively correlated with thinking disposition indicating that more analytic participants demonstrated a higher level of conflict detection. This correlation was significant when base-rates were moderate ($r = .28$, $p = .03$, $N = 60$) but not when base-rates were extreme ($r = .20$, $p = .139$, $N = 57$), although this difference between correlations was not significant

[15] The RT difference between stereotypical responses to incongruent problems and congruent problems (i.e., conflict detection) was only marginally different from 0 for moderate base-rates in the moderate first condition, RT (outliers removed): $t(30) = 1.92$, $SE = 51.3$, $p = .064$, $d = .35$; log RT (no outliers removed): $t(31) = 1.87$, $SE = .04$, $p = .07$, $d = .33$.

Table 7
Mean response time difference (in milliseconds) as a function of problem type and condition for Experiment 3. Conflict detection refers to the difference in RT between incongruent stereotypical responses and congruent items. Cognitive decoupling refers to the difference in RT between incongruent base-rate responses and congruent items.

	Conflict detection		Cognitive decoupling	
	Moderate first	Extreme first	Moderate first	Extreme first
Moderate base-rates	86 (57)	121 (58)	306 (96)	310 (78)
Extreme base-rates	574 (114)	197 (116)	284 (106)	409 (108)

Note: Standard error is listed in brackets.

Table 8
Pearson product–moment correlations between thinking disposition score and RT difference scores. Thinking disposition = sum of the Actively Open-minded Thinking scale and the Need for Cognition scale. Conflict detection refers to the difference in RT between incongruent stereotypical responses and congruent items. Cognitive decoupling refers to the difference in RT between incongruent base-rate responses and congruent items. Coefficients in bold are significant, $p < .05$. *N*'s vary.

	1	2	3	4	5	6	7
1. Thinking disposition	–	**.26**	−.16	**.28**	−.18	.20	−.11
2. Conflict detection (overall)		–	**−.30**	**.87**	**−.26**	**.72**	−.24
3. Cognitive decoupling (overall)			–	−.23	**.72**	**−.30**	**.89**
4. Conflict detection (moderate)				–	−.21	**.35**	−.21
5. Cognitive decoupling (moderate)					–	−.19	**.50**
6. Conflict detection (extreme)						–	−.23
7. Cognitive decoupling (extreme)							–

by a William's test ($t = 0.54$, $p = .59$). Future research could investigate whether differences in thinking disposition are more important under conditions where the conflict between cognitive outputs is less salient. Finally, log RT difference scores for base-rate responses did not correlate with thinking disposition, although the correlation coefficients were negative.

4.3. Discussion

Participants who, based on a self-report measure of thinking disposition, are more likely to think analytically had a larger increase in RT for stereotypical responses to incongruent relative to congruent problems. This replicates previous work (Pennycook et al., 2014a) and extends the results of Experiments 1 and 2, thereby suggesting that individual differences play a role in the responsiveness to conflict. On its own, however, this finding does not rule out the possibility that analytic individuals simply spend more time thinking when given reasoning problems. There are three additional observations that support the conclusion that individual differences in cognitive style increase the responsiveness to cognitive conflicts *in particular*. First, the reported correlations were between thinking disposition and the *difference* between RT for incongruent and congruent problems. Thus, in essence, variation in the amount of time spent on congruent problems was controlled for in the analysis. This means that the reported association is specific to variation in RT for incongruent problems (indeed, thinking disposition did not correlate with RT for congruent problems, $r < .01$, $p = .959$). Second, thinking disposition was not correlated with RT for base-rate responses to incongruent problems (in fact, it was nominally negatively correlated, see Table 8). Thus, if anything, more analytic participants took *less* time to decouple from the intuitive stereotype. Third, the association between thinking disposition and RT for stereotypical responses was somewhat more apparent when base-rates were moderate relative to extreme. This is consistent with the hypothesis that individual differences in cognitive style should become more important in cases where the conflict is more difficult to detect.

In Experiment 2, we successfully decreased the difference in RT between stereotypical responses for incongruent problems relative to congruent problems by changing the base-rates from extreme (e.g., 995/5) to moderate (e.g., 700/300); an effect that was reported across experiments in earlier work (Pennycook, Fugelsang, et al., 2012) and that replicated in Experiment 3. Moreover, in

Experiment 3, we manipulated the order of presentation for moderate and extreme base-rates and, as predicted, this had an effect on the RT increase for stereotypical responses to incongruent relative to congruent problems. Specifically, presenting participants with extreme base-rates prior to moderate ones led to an increased RT difference for moderate base-rates, suggesting that earlier conflict detection for extreme base-rates facilitated later conflict detection sensitivity for moderate base-rates. Moreover, it appears that switching the base-rates from moderate to extreme led to a large increase in the RT difference for stereotypical responses, suggesting that the already salient extreme base-rates became even more salient when presented after moderate base-rates. This indicates that rationalization following successful conflict detection is not even at ceiling when given a large number of extreme base-rates under standard conditions.

5. Experiment 4

Thus far we have successfully diminished the RT increase for stereotypical responses that is thought to index conflict detection by manipulating the extremity (Experiments 2 and 3) and order (Experiment 3) of base-rate presentation. As predicted, these manipulations had no effect on the RT increase for base-rates responses that is thought to be reflective of increased Type 2 processing due to cognitive decoupling. Moreover, individual differences in thinking disposition were positively correlated with RT for stereotypical responses but uncorrelated with RT for base-rate responses. These findings indicate a clear dissociation between conflict detection and cognitive decoupling as alternative sources of analytic engagement. However, a stronger test would be to find a manipulation that increases one and decreases the other (or vice versa). This would be very compelling evidence for the functional independence of cognitive decoupling on one hand and conflict detection on the other.

Previous research has shown that presenting base-rates *after* stereotypical information increases the likelihood and degree of base-rate use (Krosnick, Li, & Lehman, 1990). It is possible that including the base-rates as the last piece of information just prior to judgment (as was done in Experiments 1–3) increased conflict detection responsiveness. More specifically, if we assume that a given piece of information is at its most salient just prior to a decision point, presenting the base-rate just prior to the judgment in the rapid-response task should maximize the likelihood of recognition of a conflict with the previously presented stereotypes. Under this logic, presenting the base-rates before the stereotypes should make the conflict less salient, leading to a *smaller* RT difference between stereotypical responses to incongruent problems and congruent problems relative to when the base-rates are presented second. In terms of the three-stage model, this manipulation can be seen as an attempt to maximize the probability that both sources of Type 1 outputs (i.e., both IR_1 and IR_2) will enter the conflict monitoring module at similar times. With respect to cognitive decoupling, on the other hand, receiving the stereotypes after the base-rates should make them even harder to override, leading to a *larger* RT difference between base-rate responses to incongruent problems and congruent problems relative to when stereotypes are presented first (as in Experiments 1–3). In other words, successful cognitive decoupling (i.e., responding IR_2) will take more time if IR_1 is more salient.

5.1. Method

5.1.1. Participants

Eight-eight University of Waterloo undergraduates volunteered to take part in the study in return for partial course credit (23 male, 65 female, $M_{age} = 19.8$, $SD_{age} = 1.7$). Participants were randomly assigned to one of four conditions based on two between subject manipulations: (1) Order: stereotype first or base-rate first; (2) extremity: moderate or extreme base-rates. Given that the previous three experiments were run in the same participant pool, we added a question at the end of the experiment asking participants if they had seen similar problems before in previous studies (this included classic versions of base-rate problems that have also been included in multiple studies). In total, 17 participants answered this question affirmatively. However, as none of the subsequent analyses were meaningfully changed when they were excluded (apart from that which would be expected given the decrease in sample size) we retained the full sample of participants. These data were not analyzed

until the full sample was completed. All dependent variables relevant to our target research questions that were analyzed for this experiment are reported below. All manipulations are reported in the method section.

5.1.2. Materials and procedure

The materials and procedure for the rapid-response task were identical to Experiment 2 with the exception that order of base-rates/stereotypes was manipulated between subjects.

5.2. Results

5.2.1. Choice proportion for high base-rate alternative

All participants chose the response consistent with both base-rate and stereotype for congruent problems more than 80% of the time. We entered the proportion of base-rate responses into a 2 (Congruency: incongruent, congruent) × 2 (Extremity: moderate, extreme) × 2 (Order: stereotype first, base-rate first) mixed ANOVA (Table 9). There was a decrease in proportion of base-rate responses for incongruent relative to congruent items, $F(1,84) = 257.60$, $MSE = .05$, $p < .001$, $\eta^2 = .75$. As in Experiments 2 and 3, the proportion of base-rate responses was lower for moderate base-rates relative to extreme base-rates, $F(1,84) = 4.16$, $MSE = .06$, $p = .045$, $\eta^2 = .05$, and there was an interaction between congruency and extremity, $F(1,84) = 5.43$, $MSE = .05$, $p = .022$, $\eta^2 = .06$, indicating that the difference between moderate and extreme base-rates was larger for incongruent relative to congruent items. There was no three-way interaction between congruency, extremity, and order, $F < 1$.

Order had an effect on the overall proportion of base-rate choices, $F(1,84) = 10.72$, $MSE = .06$, $p = .002$, $\eta^2 = .11$, indicating more base-rate selections when base-rates were presented after stereotypes relative to when they were presented before. There was also a congruency by order interaction, $F(1,84) = 11.56$, $MSE = .05$, $p = .001$, $\eta^2 = .12$. As with extremity, this indicates that the effect of order was primarily evident among incongruent items (see Table 9). Overall, the effect of order on performance for incongruent problems was quite striking. Presenting the base-rates prior to the stereotypes led to a 28% decrease in base-rate responses in the moderate condition and a 21% decrease in the extreme condition.

5.2.2. Response time

We analyzed both raw response times (RTs) and RTs following a conversion to $\log^{10}$. Outlying raw RT's (3+ *SD*) were excluded prior to our calculation of cell means, representing 0.8% of the data. Dependent variables were entered into a 2 (Extremity: moderate, extreme) × 2 (Order: stereotype first, base-rate first) × 3 (Congruency: responses consistent with base-rate/stereotype for congruent items, responses consistent with base-rate for incongruent items, responses consistent with stereotype for incongruent items) mixed design ANOVA. Seven participants were not included in the ANOVA because they did not give any stereotypical responses. Mean RTs can be found in Table 10. Descriptive statistics for all dependent variables can be found in Supplementary materials (Tables S8–S14).

There was a main effect of congruency on RT, $F(1.5,110.7) = 17.35$, $MSE = 997027.6$, $p < .001$, $\eta^2 = .19$[4] (see Table 10) and log RT, $F(1.3,103.6) = 37.91$, $MSE = .07$, $p < .001$, $\eta^2 = .33$. There was no between-subject effect of presentation order for RT, $F(1,76) = 1.57$, $MSE = 1127209.2$, $p = .214$,

Table 9
Mean choice proportion for high base-rate alternative as a function of problem type and condition for Experiment 4.

	Stereotypes first		Base-rates first	
	Congruent	Incongruent	Congruent	Incongruent
Moderate	.97 (.03)	.46 (.38)	.95 (.03)	.18 (.21)
Extreme	.96 (.05)	.58 (.31)	.96 (.05)	.37 (.43)

Note: Standard deviations are listed in brackets.

Table 10
Mean response time (in milliseconds) as a function of problem type, response (either consistent with stereotype or base-rate), and condition for Experiment 4.

Presentation order	Base-rate extremity	Congruent	Incongruent	
			Stereotypical	Base-rate
Stereotype first	Moderate	683 (81)	944 (229)	1419 (264)
Stereotype first	Extreme	809 (81)	1781 (229)	1268 (264)
Base-rate first	Moderate	959 (79)	1179 (224)	1715 (258)
Base-rate first	Extreme	844 (83)	1176 (235)	2062 (271)

Note: Standard error is listed in brackets.

η^2 = .03, but there was an effect for log RT, $F(1,77)$ = 6.02, MSE = .13, p = .016, η^2 = .07. However, there was an interaction between presentation order and congruency for both RT, $F(2,152)$ = 3.68, MSE = 726198.9, p = .028, η^2 = .05, and log RT, $F(2,154)$ = 5.21, MSE = .04, p = .007, η^2 = .06. In addition, there was a significant three-way interaction between presentation order, base-rate extremity, and congruency for RT, $F(2,152)$ = 3.09, MSE = 726198.9, p = .048, η^2 = .04, though it was not significant for log RT, $F(2,154)$ = 1.12, MSE = .04, p = .331, η^2 = .01. All other analyses did not reach significance, all F's < 1.88, all p's > .17.

To understand the interacting effects of presentation order, we computed the two RT difference scores (as in Experiments 1–3): i.e., the difference between RTs for incongruent stereotypical and congruent (conflict detection), and the difference between RTs for incongruent base-rate and congruent (cognitive decoupling; see Fig. 5). We then ran a mixed ANOVA with response type (incongruent stereotype, incongruent base-rate) as a within-subject variable and both presentation order and base-rate extremity as between-subject variables. There was a main effect of response type, RT: $F(1,76)$ = 4.10, MSE = 1165181.7, p = .046, η^2 = .05, log RT: $F(1,77)$ = 12.26, MSE = .08, p = .001, η^2 = .14, indicating a larger overall RT difference for base-rate responses (M = 792 ms) than stereotypical responses (M = 446 ms). Crucially, there was an interaction between response type and presentation order, RT: $F(1,76)$ = 4.58, MSE = 1165181.7, p = .036, η^2 = .06, log RT: $F(1,77)$ = 5.96, MSE = .08, p = .017, η^2 = .07. There was also a marginal three-way interaction for RT, $F(1,76)$ = 4.58, MSE = 1165181.7, p = .054, η^2 = .05, however, it was not significant for log RT, $F(1,77)$ = 1.32, MSE = .08, p = .254, η^2 = .02, and therefore will not be further considered. No other effects were significant, all F's < 3.0, p's $\geqslant$ .09.

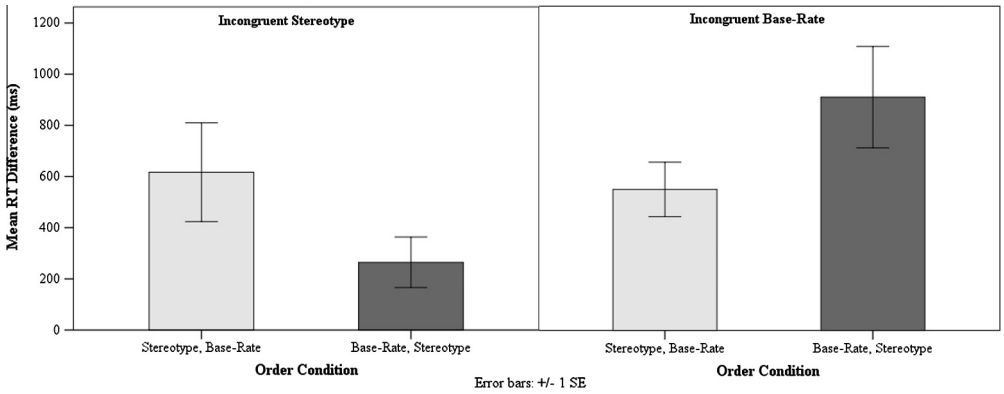

Fig. 5. Mean response time differences (in milliseconds) as a function of presentation order condition (i.e., stereotypes first or base-rates first) for Experiment 4. Incongruent stereotype refers to the difference in RT between incongruent stereotypical responses and congruent items – the conflict detection effect. Incongruent base-rate refers to the difference in RT between incongruent base-rate responses and congruent items – the cognitive decoupling effect.

The interaction between presentation order and response type is depicted in Fig. 5. Although all effects were greater than zero, all t's > 2.6, p's $\leqslant$ .011, indicating successful conflict detection and an RT increase due to cognitive decoupling in both presentation order conditions, as predicted, the order in which base-rates and stereotypes were presented had opposing effects on the RT differences for stereotype and base-rate responses. Whereas presenting the base-rate *after* the stereotype nominally increased RT for stereotypical responses (indicating an increase in conflict detection responsiveness when the less salient base-rate information is presented just before judgment), presenting the base-rate *before* the stereotype nominally increased RT for base-rate responses (indicating that presenting the intuitive stereotype just prior to judgment made it more difficult to inhibit). However, despite the significant interaction, the between-subject comparison between presentation order conditions was largely non-significant: incongruent stereotype RT, $F(1,77) = 2.73$, $MSE = 893495.5$, $p = .103$, $\eta^2 = .03$; incongruent stereotype log RT, $F(1,77) = 2.58$, $MSE = .05$, $p = .112$, $\eta^2 = .03$, incongruent base-rate RT, $F(1,83) = 2.54$, $MSE = 1085253.7$, $p = .115$, $\eta^2 = .03$; incongruent base-rate log RT, $F(1,84) = 5.18$, $MSE = .08$, $p = .025$, $\eta^2 = .05$. We were, on the other hand, able to successfully replicate the between-subject difference between extreme and moderate base-rates for incongruent stereotype RT, $F(1,77) = 3.95$, $MSE = 893495.5$, $p = .05$, $\eta^2 = .05$ and log RT, $F(1,77) = 4.25$, $MSE = .05$, $p = .043$, $\eta^2 = .05$. As in Experiments 2 and 3, base-rate extremity had no effect on RT for base-rate responses, F's < 1.

5.3. Discussion

Our three-stage model predicts that conflict detection and cognitive decoupling are distinct and separable sources of analytic engagement. In our final experiment we were able to doubly dissociate the RT increases that result from conflict detection and cognitive decoupling by simply manipulating the presentation order of base-rate and stereotype information. This, along with the differential effects of base-rate extremity (Experiments 2 and 3) and individual differences (Experiment 3) reported earlier, represents strong evidence for these distinct sources of analytic engagement.

6. General discussion

Bias is one of the most striking features of human cognition. The human mind evidently has immense intellectual capabilities – science and technology have, for example, been used to bring us to the moon and effectively abolish a great number of diseases. Given the achievements of the human race, it is perhaps reasonable to question the idea that our cognitive architecture is faulty in a fundamental way. And yet, despite the achievements, bias and irrationality also seem to confront us at every turn. People believe that the moon landing was an elaborate hoax and argue, rather dangerously, that vaccines lead to autism (Lewandowsky, Oberauer, & Gignac, 2013). 'Truthiness' often seems to be as prevalent as truth. It seems self evident that, even with humanities great achievements, there is great value in determining the factors that underlie our biases. Our goal with this line of studies was to speak to this question by investigating and elucidating integral features of human cognitive architecture. We have introduced a three-stage model of analytic engagement and, using experimental manipulation and individual differences, we have dissociated two of the integral components of the model: conflict detection and cognitive decoupling. Below we will further explicate the model and compare it to other perspectives by drawing on current and past data.

6.1. The three-stage model of analytic engagement: summary of current evidence

Our experiments speak to the utility of our three-stage model in three fundamental ways: (1) We dissociated increases in Type 2 processing that indicate, on one hand, rationalization following successful conflict detection and, on the other hand, cognitive decoupling, (2) we found that conflict monitoring sometimes fails in predictable ways, and (3) we demonstrated that individual differences modulated conflict detection responsiveness. Specifically, making the base-rates moderate (e.g., 700 nurses and 300 doctors) instead of extreme (e.g., 995 nurses and 5 doctors) selectively decreased

the conflict detection effect. This was evident both between subjects (Experiment 2) and within subjects (Experiment 3). The conflict detection effect, but not the cognitive decoupling effect, was also selectively affected by the order of extreme versus moderate base-rates in Experiment 3. Finally, changing the order of base-rate/stereotype presentation evidenced a double dissociation between conflict detection and cognitive decoupling. Namely, presenting base-rates prior to stereotypes led to a decrease in the conflict detection effect and an increase in the cognitive decoupling effect relative to the opposite orientation.

In Experiments 1 and 2, the cases where participants did not appear to detect the conflict (i.e., a negative difference in RT between stereotypical responses to incongruent versus congruent problems) were limited to particularly biased participants. This indicates that some participants failed to detect the conflict between base-rates and stereotypes. Consistent with previous research (Pennycook et al., 2014a), analytic thinking disposition was positively correlated with the conflict detection effect and (nominally, but not significantly) negatively correlated with the cognitive decoupling effect – evidently, individual differences in active open-mindedness is more consequential for increases in analytic processing attributable to rationalization following successful conflict detection than increases in analytic processing attributable to the active suppression of an intuitive response.

6.2. Conflict detection failures

Although the distinction between conflict detection and cognitive decoupling has not, until now, been formally built into a dual-process model, it is also not inconsistent with any current models. A more controversial dimension of our three-stage model relates to the potential for conflict detection failures. According to our model, it is possible for a conflict between two intuitive outputs to be present and not detected. Further, we have claimed that this may occur in cases where a second initial response (IR_2) comes to mind much less quickly and fluently than a first initial response (IR_1). This account contrasts with earlier work where it has been claimed that conflict detection is highly efficient (e.g., De Neys, 2012, 2014).

Pennycook, Fugelsang, et al. (2012) provided the first evidence that conflict detection may not be as efficient as previously claimed. Specifically, there was no evidence of an overall conflict detection effect (as indexed by RT) when base-rates were made moderate instead of extreme. In Experiments 2–4 in the current manuscript, we employed this manipulation again but instead found evidence for conflict detection given moderate base-rates across each of the experiments. The conflict detection effect was larger for extreme than moderate base-rates, suggesting that participants were more *responsive* to conflict in the former case, but a question remains: Is there direct evidence for *categorical* failures of conflict detection?

Following a recent analysis by Mevel et al. (2015), we isolated the proportion of participants who had actually took *longer* to respond to congruent problems than to give stereotypical responses to incongruent problems. By this very conservative analysis, 17.2–17.9% of the participants in Experiments 1 and 2 showed no evidence of detecting the conflict when given extreme base-rates. In Experiment 2, 28.6% of the participants in the moderate base-rate condition had a negative difference between congruent and incongruent problems. Importantly, these ostensible cases of categorical conflict detection failures were also associated with very low rates of base-rate responding (ranging from 5.8% to 11% against overall means of 28–49%). This indicates that these cases should not be attributed to random sampling error but are rather representative of a group of participants who, potentially, are highly biased precisely *because* they failed to detect the conflict between base-rates and stereotypes. These results clearly illustrate that conflict detection is not perfectly efficient and, as a consequence, that detection failures are one source of biases in reasoning.

Our three-stage model highlights two potential sources of detection failures: (1) no second (conflicting) initial response (IR_2) is elicited, or (2) a second initial response is elicited, but the probability of conflict detection is dependent on the relative speed at which the competing initial responses come to mind. Although distinguishing between these possibilities empirically is outside the scope of this paper, it is worthwhile discussing how they can be accommodated in the model. Consider the difference between extreme and moderate base-rates, for example. If the first potential source is the explanation of the apparent detection failure reported by Pennycook, Fugelsang, et al. (2012), then extreme

base-rates must cue initial responses and moderate base-rates must not. This would presume that the Type 1 process involved in autonomously recognizing and responding to base-rates is specific to extreme cases. This seems rather unlikely because the knowledge required for the base-rates to enter into judgment is (a) rudimentary, with the ability to process probabilities having potential origins in childhood (Denison, Reed, & Xu, 2013; Denison & Xu, 2014; Xu & Denison, 2009) and (b) equivalent regardless of base-rate extremity (i.e., the same normative standards hold for extreme and moderate base-rates). An alternative possibility is that the Type 1 process that responds to base-rates is less specific, but the speed at which the Type 1 outputs come to mind is nonetheless dependent on the extremity of the base-rate. A less efficient conflict monitoring system could be influenced by such a factor. Clearly, further work is necessary to better understand the nature of conflict monitoring in reasoning. This is one way in which the field could benefit from increased discussion of formal models of conflict detection and analytic engagement.

6.3. Dual process theories and the problem of bias

Our three-stage model accommodates a more nuanced approach to the problem of bias than previous perspectives because it allows for both failures of analytic engagement and failures of response inhibition. These mechanisms are associated with two different – and often competing – dual-process explanations for the pervasiveness of bias in reasoning and decision making. The traditional dual-process view is that bias primarily results from a failure to sufficiently engage analytic reasoning mechanisms that might be used to override intuitive responses. This view is typically associated with Evans' default-interventionist model which emphasizes the need for Type 2 processing to intervene against a default intuitive response (e.g., Evans, 2007; see also Stanovich, 2009a). Thompson's metacognitive model also fits into this category as well, as it highlights the role of salient feelings of rightness that pre-empt Type 2 processing (e.g., Thompson, 2009). These models assume that humans often fail to detect the need to engage the very processing that could potentially undermine bias (see De Neys, 2014 for further discussion).

In contrast, De Neys suggests that bias results primarily from inhibition failures (e.g., De Neys, 2012). According to this less traditional view, participants successfully engage Type 2 processing when the problem contains some sort of response conflict (i.e., they succeed at conflict detection), but simply fail to do so effectively (i.e., they fail at cognitive decoupling). Although De Neys (2014) has discussed potential boundary conditions of conflict detection, the discussed examples all included cases where the problem fails to cue a logical intuition. For example, the abstract Wason card selection task may not cue competing intuitive responses given its very low accuracy rates (see Wason & Evans, 1975). Participants may not develop the requisite knowledge to be able to solve such complex problems easily (i.e., barring an intuitive lure). However, such problems do not speak to the issue of conflict monitoring *efficiency* because, based on the logical intuition model, if no logical intuitions are elicited by the problem there is simply no conflict to detect.

As clarified recently by De Neys (2014), a key question for this debate has to do with the modal biased reasoner (see also, Mevel et al., 2015). It may be that bias arises sometimes from failures of analytic engagement (Evans, 2007; Kahneman, 2003) and sometimes from inhibition failures (De Neys, 2012), perhaps depending on contextual or individual difference factors. The operative question, then, is which type of failure is more common? As discussed, failures of analytic engagement may be particularly influential given complex reasoning problems such as the Wason card selection task because the probability of competing intuitions is low. In the context of less complex reasoning and decision making problems, such as the one employed in this investigation, the bias exhibited by the modal reasoner appears to be more the result of inhibition failures (De Neys, 2012) than of outright failures of analytic engagement (Evans, 2007). This is consistent with the wealth of evidence for successful conflict detection over a wide range of heuristic and biases tasks (see De Neys, 2012, 2014). Indeed, these results appear to be robust even when the analysis is isolated to the first item presented (e.g., De Neys, Rossi, & Houdé, 2013). One weakness of the literature, however, is that most investigations of conflict detection have focused entirely on the presence or absence of a significant *overall* conflict detection effect and therefore have not isolated the prevalence of *categorical* failures across participants (Mevel et al., 2015 being the exception). This is an important area for future research.

Although the debate about the modal biased reasoner is important in the context of individual reasoning paradigms, we hasten to add that these perspectives are not mutually exclusive and, in fact, our three-stage model accommodates both (see also, Stanovich & West, 2008; Fig. 1). We even found evidence for both detection *and* inhibition failures using the rapid-response base-rate task. Here, participants took longer on conflict problems relative to the non-conflict baseline problems when the analysis was isolated to cases when stereotypical ('biased') responses were given. This indicates that they were able to detect the conflict but failed to inhibit the intuitive stereotypical response. However, we also found evidence for large variability in this conflict detection effect – both across individuals and as a result of subtle manipulations. Indeed, a number of the relatively more biased individuals were not apparently able to detect the conflict between moderate base-rates and salient stereotypes. This indicates that detection failures also occur. These findings are easily accommodated by our three-stage model of analytic engagement.

6.4. Dual-processing: theory, metatheory, and criticisms

Although dual-process theory is widely accepted and has been applied in multiple domains of psychology (see Evans, 2008; Evans & Stanovich, 2013a), there have been a number of recent critiques (e.g., Keren, 2013; Keren & Schul, 2009; Kruglanski, 2013; Kruglanski & Gigerenzer, 2011; Osman, 2004, 2013). There are two general classes of criticism that are levied against dual-process theories: (1) Sufficient evidence for two types of processes is lacking and the extant data can just as easily, and more parsimoniously, be explained by unimodal theories, and (2) dual-process theories are unfalsifiable, are poorly defined, fail to motivate new questions and yield testable predictions, or some combination of these things. The current work represents a consequential implementation of a dual-process perspective and, as such, it is worthwhile discussing how it speaks to such theoretical and metatheoretical debates. It is also necessary to situate our own three-stage dual-process model within the context of these criticisms.

Evans and Stanovich (2013a) have argued that Type 1 and Type 2 processes are qualitatively different because Type 1 processes are autonomous (i.e., "... the execution of Type 1 processing is mandatory when their triggering stimuli are encountered..." p. 236) whereas Type 2 processes require a deliberative instantiation of working memory resources. Kruglanski (2013, see also Kruglanski & Gigerenzer, 2011) has argued that because the speed at which something autonomously comes to mind (via Type 1 processing) is likely dependent on the strength of the stimulus–response pairing, the responses that either come to mind or are generated later in the reasoning process are simply those that are associated with a weaker stimulus–response pairing (for a related argument, see Osman, 2004, 2013). In other words, Type 1 and 2 outputs differ based on a continuum. However, consider the following argument (Kruglanski, 2013): "If the quickly activated thought [a Type 1 intuition, or IR_1 in our model] seemed appropriate to the cognizer's task, it might be adopted and acted upon. If it seemed less than satisfactory, the individual may keep on searching for more appropriate albeit less accessible notions, but only if she had the motivation and mental resources to do so (see Kruglanski et al., 2012)" (p. 249). Here the argument appears to fall back to the familiar dual-process dichotomy. The initial process that generates outputs based on stimulus–response pairings is supplemented by a later process that determines if the initial output is satisfactory and that may initiate a search for alternatives. This is captured in our three-stage model, though we (unlike Kruglanski) emphasize the distinction between the initial and later processes. This distinction – or, in other words, our dual-process perspective – allows us to not only explain why thoughts or responses are generated, but why they might be considered further. If the initial generated response (IR_1) does not conflict with an additional autonomously generated response (IR_2), this further consideration may be cursory. However, when a conflict is successfully detected, the reasoner will spend more time and effort thinking analytically. This may take the form of rationalization or cognitive decoupling, the latter of which allows for novel combinations of distant semantic concepts (see Barr et al., 2014) – a process that is difficult to accommodate in a rule-based stimulus–response model.

The second major criticism of dual-process theories relates to the difference between theory and metatheory (Evans & Stanovich, 2013b). Dual-process theories in their general form are metatheoretical in that they distinguish between 'intuitive' and 'reflective' types of processes but do not elaborate

on how that distinction bears on any given task. As a consequence, general dual-process theories are not falsifiable and do not lead to testable predictions (see Keren, 2013; Keren & Schul, 2009). However, a dual-process perspective can be used to generate testable and falsifiable models that are specific to a type of task or phenomenon. If done successfully, this evidences the *value* of dual-process theories as a metatheoretical perspective. Moreover, the likelihood that the distinction between intuition and reflection is misguided becomes decreasingly small with increasing numbers of successful applications of dual-process theory.

Our three-stage model represents a testable and falsifiable instantiation of dual-process theory. The model is specific in that it was designed to explain the cognitive processes involved in solving the types of tasks or problems that contain conflicting sources of information, but general in the sense that conflicting information is presumably very common. Moreover, the model is clearly defined and makes straightforward predictions. We consider the current work to be a representative case for a meaningful and successful application of a dual-process metatheoretical perspective.

6.5. Beyond base-rate neglect

Although we focused entirely on base-rate problems here, the three-stage model should be applicable in any case where a problem or cue may engender conflicting responses. To illustrate this point we will outline two further examples. The first will serve as an example of how the three-stage model accommodates recent experimentation in a traditional reasoning paradigm (namely, belief bias in syllogistic and conditional reasoning). The second example – goal conflict – will illustrate how the three-stage model can be applied to a different area of research altogether.

6.5.1. Belief bias

Belief bias refers to the tendency to endorse the conclusion of a deductive argument based on its believability instead of its logical structure. Consider the following example (Sá, West, & Stanovich, 1999):

All plants need water.
Roses need water.
Therefore, roses are plants.

This syllogism contains a conflict between logic and belief such that the conclusion is logically invalid (i.e., the conclusion does not follow from the premises) but nonetheless believable (i.e., roses are indeed plants). As a result, participants often incorrectly endorse the conclusion as logically valid (Evans et al., 1983). Nonetheless, a number of studies have demonstrated that participants are able to detect the conflict between logic and belief; a finding that applies to both syllogisms (Ball, Philips, Wade, & Quayle, 2006; De Neys & Franssens, 2009; De Neys, Moyens, et al., 2010; Stupple & Ball, 2008; Stupple, Ball, & Ellis, 2013) and conditionals (Handley et al., 2011). According to the three-stage model (see also, De Neys, 2012), this conflict detection indicates that some or most participants must be intuitive logicians. For a conflict between logic and belief to be reliably detected, both factors must cue a Type 1 output. Presumably, then, belief bias tends to dominate logic partially because belief provides a quicker, more salient Type 1 output.

There is evidence to support the counterintuitive claim that all logic does not necessarily require Type 2 processing. In a typical syllogistic reasoning study, participants are informed to assume that all premises are true and that they should only endorse a conclusion if it necessarily follows from the premises. Under these instructions, logical responding decreases when the influence of Type 2 processing is diminished through a time deadline (Evans & Curtis-Holmes, 2005) or secondary task (De Neys, 2006). These findings, along with the positive correlation between cognitive capacity and logical responding (Sá et al., 1999), indicate that logical reasoning *requires* Type 2 processing. However, across a number of experiments, Handley et al. (2011) employed an instruction manipulation where participants were asked to give a logical response (as in previous studies) or belief-based response. Contrary to what would be expected if logic *requires* Type 2 processing, participants took longer for conflict than non-conflict problems *regardless of the instruction manipulation*. In other words, the logical structure of

the problems interfered with belief-based responses. This cross-interference effect replicated across different presentation formats and using both conditionals and syllogisms. The claim that belief and logic may cue a Type 1-Type 1 conflict is, at the very least, made plausible by these findings (see Handley & Trippas, 2015).

6.5.2. Goal conflict

Cases where one's desire trumps a potentially more beneficial goal are common and, as such, have been the focus of much philosophical and psychological debate (e.g., Baumeister, Heatherton, & Tice, 1994; Mele, 1995; Muraven & Baumeister, 2000; Thaler & Shefrin, 1981). A great deal of psychological research has highlighted how self-control can be accomplished through an override of salient desires or impulses via finite cognitive resources (e.g., Baumeister, Bratslavsky, Muraven, & Tice, 1998; Hagger, Wood, Stiff, & Chatzisarantis, 2010; Vohs & Heatherton, 2000). However, more recent research has highlighted an alternative source of self-control lapses: i.e., the failure to recognize the conflict between a desire and a goal in the first place (e.g., Hofmann & Kotabe, 2012; Myrseth & Fishbach, 2009).

According to the three-stage model, the probability of detecting a conflict between a standard (goal) and a temptation/desire will be determined by the relative speed at which the two (or more) representations come to mind. In other words, the key to conflict detection in the case of self-control conflicts is the degree to which a standard (IR_2) lags a temptation/desire (IR_1). Consider the case of a New Year's resolution to eat healthier. This is a relatively salient goal on January 1st but less salient on February 1st. If, on the 1st of January, a piece of cake is encountered, it is likely that the conflict will be detected because the goal should come to mind fast enough to interrupt the decision to eat cake. On February 1st, in contrast, the goal still exists but is clearly less salient. In such cases it is less likely that the individual will recognize that they are about to commit a self-control failure (see Karlan, McConnell, Mullainathan, & Zinman, 2010). This pattern could also be influenced by the salience of the temptation. It would be difficult even on January 1st for a very hungry person to recognize that eating a slice of cake conflicts with a more abstract goal. Although this is clearly speculative, this example demonstrates how the three-stage model of analytic engagement could be applied in an entirely different domain of research.

6.6. Limitations and further specifications

6.6.1. Other sources of Type 2 processing

The goal of the three-stage model is to elucidate the low-level cognitive sources of analytic engagement, and we have highlighted conflict monitoring as a key mechanism. As such, we have used a paradigm where participants think more analytically about some items relative to others in the absence of an explicit cue such as an instruction manipulation. This should not be taken to imply, however, that analytic processing *cannot* occur in the absence of successful conflict detection. Explicit cues to think analytically (such as an instruction to think logically, see Daniel & Klaczynski, 2006; Evans et al., 1994, 2010; Vadeboncoeur & Markovits, 1999) may alter the course of analytic reasoning after conflict monitoring has been completed – for example, by shifting a participant from simply verifying an initial response to rationalizing or even decoupling. Indeed, an added benefit of the three-stage model is that it permits a higher degree of specificity when discussing alternative sources of Type 2 processing. For example, it is possible for other types of interventions, such as practice or learning effects, to affect earlier stages of the reasoning process by altering the probability and speed at which initial responses come to mind (which, in turn, may affect the probability of analytic engagement following successful conflict detection).

6.6.2. Other classes of Type 2 processing?

According to the three-stage model, decoupling and rationalization may be considered different "classes" of Type 2 processes. If analytic thought is engaged for more than simply verifying the initial response as adequate, the reasoner must either focus thought on the initial response (rationalization), suppress it in lieu of some other output (decoupling), or do some combination of these (i.e., moving back and forth between the two over time). However, it is important to note that there are naturally

many forms that such processing takes. Consider, for example, a stereotype wherein a person described as shy and nerdy is initially judged to be a computer technician. One might rationalize this initial response by engaging hypothetical thought to simulate a computer technician conference full of shy and nerdy people. Or, perhaps, one might decouple from this stereotype by suppressing it and engaging hypothetical thought to imagine an outgoing computer technician. In this sense, decoupling and rationalization describe the association between whatever processing is occurring and the initial response.

The three-stage model is consistent with default-interventionist models and may even be considered a default-interventionist model itself because Type 2 processing does not occur until *after* Type 1 processes output a response. The primary difference between the three-stage model and traditional default-interventionist models (e.g., Evans, 2007, 2010a, 2010b) is that the former is interested in the *causes* of analytic intervention whereas the latter are typically focused on determining the common defaults that undermine reasoning (e.g., prior beliefs) and the problem factors that require intervention to enter into reasoning (e.g., logical validity). In other words, our three-stage model is focused on the "how" and previous default-interventionist models are typically focused on the "what". In the three-stage model, it is possible for a factor traditionally associated with analytic processing such as base-rate probabilities or logical validity to be the source of a Type 1 output (see Handley & Trippas, 2015) – and, in fact, for some individuals it is quite possible that factors such as logic are more intuitive than factors such as belief (that is, logic cues IR_1 and belief cues IR_2). Moreover, manipulating problem structure (e.g., the format of base-rates, Evans, Handley, Perham, Over, & Thompson, 2000) or top-down factors such as instructions (Handley et al., 2011; Pennycook, Trippas, et al., 2014) may alter the initial processing of the problems such that typically more intuitive responses are engendered more slowly (from IR_1 to IR_2) and typically less intuitive responses are engendered more quickly (from IR_2 to IR_1). In this way, the three-stage model is capable of accommodating both logical intuitions and effortful beliefs (for an extended discussion, see Handley & Trippas, 2015).

6.6.3. Other measures

An additional limitation of the current work is that we have focused entirely on response time as an indicator of increased analytic engagement. Although RT has been used in a large number of conflict detection studies (e.g., Bonner & Newell, 2010; De Neys & Franssens, 2009; De Neys & Glumicic, 2008; Handley et al., 2011; Pennycook, Fugelsang, et al., 2012; Stupple & Ball, 2008; Villejoubert, 2009), many additional measures have been used as well, including eye tracking (Ball et al., 2006), memory recall (De Neys & Glumicic, 2008; Franssens & De Neys, 2009), verbal protocols (De Neys & Glumicic, 2008), skin conductance response (De Neys, Moyens, et al., 2010), confidence (De Neys, Lubin, & Houdé, 2013, 2014; De Neys et al., 2011, 2013; Rossi, Cassotti, Agogué, & De Neys, 2013; Stupple et al., 2013; Thompson & Johnson, 2014; Thompson et al., 2011), liking ratings (Morsanyi & Handley, 2012), and neuropsychological measures like fMRI (De Neys et al., 2008) and ERP (Banks & Hope, 2014; De Neys, Novitskiy, Ramautar, & Wagemans, 2010). Additional measures could be used to test key aspects of our three-stage model. For example, confidence ratings could reveal insights into the potential role of metacognition as an additional source of analytic engagement (discussed subsequently). Further, time-sensitive online measures of conflict sensitivity such as skin conductance and ERPs could be used to investigate the time course and relative efficiency of conflict detection. Our model suggests that conflict detection occurs early in the reasoning process and depends on the speed at which competing Type 1 outputs come to mind. A further possibility is that conflict monitoring *itself* is suspect to individual differences; perhaps as a consequence of differential anterior cingulate cortex functioning (Fornito et al., 2004).

6.7. Future directions

We have focused on testing two primary claims that were derived from the three-stage model: (1) Conflict monitoring may sometimes fail, and (2) conflict detection and cognitive decoupling are separable and dissociable *sources* of Type 2 processing. However, there are other testable claims that can be derived from the model but that have not been investigated here. Although each claim is grounded

in prior theoretical and empirical work, it is necessary to be precise about the limits of the current investigation. This will hopefully guide future research.

6.7.1. Stage 1

According to the three-stage model, a stimulus may cue multiple Type 1 outputs that come to mind at different speeds. This is a key claim that has not been assessed here. We hasten to add, however, that the idea that multiple Type 1 outputs may be engendered by the same stimulus follows directly from the uncontroversial idea that Type 1 processes operate autonomously and in parallel (see Stanovich, 1999, 2004 for discussion of the "autonomous set of systems"). Moreover, the idea that some things come to mind more quickly and fluently than others is supported by decades of metacognition research (Alter & Oppenheimer, 2009; Benjamin, Bjork, & Schwartz, 1998; Schwarz, 2004; Whittlesea, Jacoby, & Girard, 1990; Whittlesea & Leboe, 2003). More directly, decreased answer fluency is associated with increased metacognitive "feelings of rightness" which, in turn, have been implicated as a source of increased Type 2 processing (Thompson et al., 2011, 2013). Further, Thompson and Johnson (2014) provide evidence that conflict detection decreases feelings of rightness, which, in turn, mediates the extent of subsequent Type 2 thinking.

Further work is required to fully integrate metacognitive considerations into the three-stage model. As an example, the model predicts that each Type 1 output is associated with a unique speed of processing and that the relation between these Type 1 outputs will partly determine what occurs later in the reasoning process. Is fluency only relative to the final response output (i.e., following Type 2 processing in Stage 3) or does each Type 1 output engender unique fluencies which, depending on their status relative to each other, determine a unique final fluency judgment? The three-stage model could facilitate future research such as this.

6.7.2. Stage 2

The three-stage model allows for conflict monitoring failures; a component supported by both current and past data (e.g., Pennycook, Fugelsang, et al., 2012). But what causes detection failures? Assuming we accept the claim that some Type 1 outputs come to mind more quickly than others, it is plausible that the probability of conflict detection success is determined by the relative differences in the speed at which competing Type 1 outputs come to mind. This possibility has not been directly tested here. Conflict detection was less efficient when base-rates were moderate (e.g., 700 lawyers, 300 engineers) than when extreme (e.g., 995 lawyers, 5 engineers) (Experiments 2–4; see also, Pennycook, Fugelsang, et al., 2012), but this difference was mitigated if extreme base-rate problems were presented in a block prior to moderate ones (Experiment 3). Manipulating conflict detection in this way supports the idea that conflict monitoring is not perfectly efficient, but the mechanisms underlying this effect are still unclear. Does prior experience with salient extreme base-rates facilitate processing of moderate base-rates? This would decrease the difference in processing speed between moderate base-rates and stereotypes and make conflict detection more likely. Alternatively, it may be that extreme base-rates draw more attention, thereby decreasing the likelihood that they will be misrepresented at the level of language comprehension (Mata, Schubert, & Ferreira, 2014). We have focused on testing the higher level distinctions made by the three-stage model (i.e., Stages 2 and 3) at the expense of these types of lower level issues (i.e., Stage 1). Future research should focus on further specification of Type 1 processing (see Thompson, 2014).

6.7.3. Stage 3

We focused on base-rate problems to which there are only two potential responses (i.e., the stereotypical response, IR_1, or the base-rate response, IR_2) and, as such, the generation of alternative responses (AR) was not investigated despite being included in the three-stage model. According to the model, it is impossible for a response to be generated that is not sourced, to some degree, by an autonomous (Type 1) response. In other words, the Type 2 processing that occurs at Stage 3 requires an earlier Type 1 output from Stage 1. Novelty arises when semantically distant representations are combined via cognitive decoupling (i.e., "reasoned connections"; see Barr et al., 2014). This alternative response generation process allows for a response to be generated that was not initially available through Type 1 processing. For example, in the Remote Associates Test (RAT) participants are asked

to generate a common associate for a set of ostensibly unrelated words (e.g., "sore, shoulder, sweat"). Correctly solving such a problem requires an insight about the common connection between the semantically distant words (e.g., "cold"). Barr et al. (2014) theorized that this connection is facilitated by activation of Type 2 processing and, consistent with this idea, found that performance on the RAT (and other related creativity tasks) is strongly correlated with measures thought to assess individual differences in analytic thinking (see also Ball & Stevens, 2009; Chein & Weisberg, 2014). This indicates that an initially unavailable response became available through an increase in analytic engagement, perhaps as a result of iterations of the processes described in the three-stage model.

7. Conclusion

What makes us think? Of interest here are not the more obvious content-related answers to this question – a good book or a stimulating conversation – rather, our goal was to better understand the cognitive architecture of analytic thought. To this end, we proposed a three-stage dual-process model that combines elements of previous reasoning models with novel insights. We also provided evidence for integral components of our model from response time analyses using a rapid-response base-rate task.

Although the question of what cues analytic thought has received some attention in recent years (e.g., De Neys, 2012; Evans, 2009; Stanovich, 2009a; Thompson, 2009), there is still a great deal of work to be done. This represents a rather striking gap in our knowledge, as the capacity to think and reason is often considered the paragon of what makes us uniquely human. Moreover, obtaining a stronger understanding of the bottom-up factors that lead to analytic thought could lead to more efficient debiasing interventions and, as a consequence, better decision-making. Our principle goal in the current work was to inspire and guide such research. In a world where 'truthiness' too often triumphs over truth, we can scarcely think of a more important academic pursuit.

Acknowledgments

We would like to thank Jonathan St. B.T. Evans, Wim De Neys, and an anonymous reviewer for their invaluable comments. Funding for this study was provided by the Natural Sciences and Engineering Research Council of Canada.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cogpsych.2015.05.001.

References

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.

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences, 8*, 170–177.

Ball, L. J., Lucas, E. J., Miles, J. N. V., & Gale, A. G. (2003). Inspection times and the selection task: What do eye-movements reveal about relevance effects? *Quarterly Journal of Experimental Psychology A, 56*, 1053–1077.

Ball, L. J., Philips, P., Wade, C. N., & Quayle, J. D. (2006). Effects of belief and logic on syllogistic reasoning: Eye-movement evidence for selective processing models. *Experimental Psychology, 53*, 77–86.

Ball, L., & Stevens, A. (2009). Evidence for a verbally-based analytic component to insight problem solving. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the Cognitive Science Society* (pp. 1060–1065). Austin, TX: Cognitive Science Society.

Banks, A. P., & Hope, C. (2014). Heuristic and analytic processes in reasoning: An event-related potential study of belief bias. *Psychophysiology, 51*, 290–297.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioural and Brain Sciences, 30*, 241–256.

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York: Cambridge University Press.

Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2014). Reasoned connections: A dual-process perspective on creative thought. *Thinking & Reasoning, 21*, 61–75.

Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2015). The brain in your pocket: Evidence that Smartphones are used to supplant thinking. *Computers in Human Behavior, 48*, 473–480.
Baumeister, R. F., Bratslavsky, M., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology, 74*, 1252–1265.
Baumeister, R. F., Heatherton, T. F., & Tice, D. M. (1994). *Losing control: How and why people fail at self-regulation*. San Diego: Academic Press.
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.
Bonner, C., & Newell, B. R. (2010). In conflict with ourselves? An investigation of heuristic and analytic processes in decision making. *Memory & Cognition, 38*, 186–196.
Brenner, L. A., Griffin, D. W., & Koehler, D. J. (2012). A case-based model of probability and pricing judgments: Biases in buying and selling uncertainty. *Management Science, 58*, 159–178.
Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Science, 4*, 215–222.
Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*, 197–253.
Chein, J. M., & Weisberg, R. W. (2014). Working memory, insight, and restructuring in verbal problems: Analysis of compound remote associate problems. *Memory & Cognition, 42*, 67–83.
Daniel, D. B., & Klaczynski, P. A. (2006). Developmental and individual differences in conditional reasoning: Effects of logic instructions and alternative antecedents. *Child Development, 77*, 339–354.
De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoned. *Psychological Science, 17*, 428–433.
De Neys, W. (2007). Nested sets and base-rate neglect: Two types of reasoning? *Behavioral and Brain Sciences, 30*, 260–261.
De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science, 7*, 28–38.
De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning, 20*, 167–187.
De Neys, W., & Bonnefon, J. F. (2013). The whys and whens of individual differences in thinking biases. *Trends in Cognitive Sciences, 17*, 172–178.
De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE, 6*, e15954.
De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition, 113*, 45–61.
De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*, 1248–1299.
De Neys, W., Lubin, A., & Houdé, O. (2013). The smart non-conserver: Preschoolers detect their number conservation errors. *Child Development Research*.
De Neys, W., Lubin, A., & Houdé, O. (2014). The smart nonconserver: Preschoolers detect their number conservation errors. *Child Development Research*.
De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, & Behavioral Neuroscience, 10*, 208–216.
De Neys, W., Novitskiy, N., Ramautar, J., & Wagemans, J. (2010). What makes a good reasoner?: Brain potentials and heuristic bias susceptibility. *Proceedings of the Annual Conference of the Cognitive Science Society, 32*, 1020–1025.
De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review, 20*, 269–273.
De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science, 19*, 483–489.
Denison, S., Reed, C., & Xu, F. (2013). The emergence of probabilistic reasoning in very young infants: Evidence from 4.5- and 6-month-olds. *Developmental Psychology, 49*, 243–249.
Denison, S., & Xu, F. (2014). The origins of probabilistic inference in human infants. *Cognition, 130*, 335–347.
Epstein, S. (1994). Integration of the cognitive and psychodynamic unconscious. *American Psychologist, 49*, 709–724.
Evans, J. S. 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 & 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. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology, 59*, 255–278.
Evans, J. St. B. T. (2010a). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry, 21*, 313–326.
Evans, J. St. B. T. (2010b). *Thinking twice: Two minds in one brain*. Oxford, England: Oxford University Press.
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. S. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning, 11*, 382–389.
Evans, J. St. B. T. (2009). How many dual process theories do we need: One, two, or many? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 33–54). City: Oxford University Press.
Evans, J. S. B. T., Handley, S. J., Neilens, H., Bacon, A. M., & Over, D. E. (2010). The influence of cognitive ability and instructional set on causal conditional inference. *Quarterly Journal of Experimental Psychology, 63*, 892–909.
Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition, 77*, 197–213.
Evans, J. St. B. T., Newstead, S. E., Allen, J. L., & Pollard, P. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology, 6*, 263–285.
Evans, J. St. B. T., & Stanovich, K. E. (2013a). Dual-process theories of higher cognition: Advancing the debate. *Perspectives in Psychological Science, 8*, 223–241.

Evans, J. St. B. T., & Stanovich, K. E. (2013b). Theory and metatheory in the study of dual processing: Reply to comments. *Perspectives in Psychological Science, 8*, 263–271.
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.
Fornito, A., Yucel, M., Wood, S., Stuart, G. W., Buchanan, J., Proffitt, T., et al (2004). Individual differences in anterior cingulate/paracingulate morphology are related to executive functions in healthy males. *Cerebral Cortex, 14*, 424–431.
Frankish, K., & Evans, J. St. B. T. (2009). The duality of mind: An historical perspective. In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 1–32). Oxford: Oxford University Press.
Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning, 15*, 105–128.
Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science, 336*, 493–496.
Gigerenzer, G., & Regier, T. (1996). How do we tell an association from a rule? Comment on Sloman (1996). *Psychological Bulletin, 119*, 23–26.
Goel, V., & Dolan, R. J. (2003). Explaining modulation of reasoning by belief. *Cognition, 87*, 11–22.
Hagger, M. S., Wood, C., Stiff, C., & Chatzisarantis, N. L. (2010). Ego depletion and the strength model of self-control: A meta-analysis. *Psychological Bulletin, 136*, 495–525.
Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 28–43.
Handley, S. J., & Trippas, D. (2015). Dual processes, knowledge, and structure: A critical evaluation of the default interventionist account of biases in reasoning and judgment. *Psychology of Learning and Motivation, 62*.
Hofmann, W., & Kotabe, H. (2012). A general model of preventive and interventive self-control. *Social and Personality Psychology Compass, 6*, 707–722.
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. (2011). *Thinking, fast and slow*. New York: Farrar, Strauss, & Giroux.
Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases* (pp. 49–81). New York: Cambridge University Press.
Kahneman, D., & Frederick, S. (2005). A model of heuristic judgement. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge Handbook of thinking and reasoning* (pp. 267–293). Cambridge, MA: Cambridge University Press.
Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.
Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 493–508). Cambridge, England: Cambridge University Press.
Karlan, D., McConnell, M., Mullainathan, S., & Zinman, J. (2010). Getting to the top of mind: How reminders increase saving. *National Bureau of Economic Research working paper no. 16205*.
Keren, G. (2013). A tale of two systems: A scientific advance or a theoretical stone soup? Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science, 8*, 257–262.
Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science, 4*, 533–550.
Krosnick, J. A., Li, F., & Lehman, D. R. (1990). Conversational conventions, order of information acquisition, and the effect of base rates and individuating information on social judgments. *Journal of Personality and Social Psychology, 59*, 1140–1152.
Kruglanski, A. W. (2013). Only one? The default interventionist perspective as a unimodel—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science, 8*, 242–247.
Kruglanski, A. W., Belanger, J., Chen, X., Kopetz, C., Pierro, A., & Mannetti, L. (2012). The energetics of motivated cognition: A force field analysis. *Psychological Review, 119*, 1–20.
Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberative judgements are based on common principles. *Psychological Review, 118*, 97–109.
Kunda, Z. (1990). The case for motivated reasoning. *Psychology Bulletin, 108*, 480–498.
Lewandowsky, S., Oberauer, K., & Gignac, G. E. (2013). NASA faked the moon landing—Therefore, (climate) science is a Hoax: An anatomy of the motivated rejection of science. *Psychological Science, 24*, 622–633.
Liang, P., Goel, V., Jia, X., & Li, K. (2014). Different neural systems contribute to semantic bias and conflict detection in the inclusion fallacy task. *Frontiers in Human Neuroscience*.
Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on Psychological Science, 4*, 390–398.
Mata, A., Schubert, A., & Ferreira, M. B. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. *Cognition, 133*, 457–463.
Mele, A. (1995). *Autonomous agents*. New York: Oxford University Press.
Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., et al (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology, 27*, 227–237.
Morsanyi, K., & Handley, S. J. (2012). Does thinking make you biased? The case of the engineers and lawyer problem. *Proceedings of the Annual Meeting of the Cognitive Science Society, 34*, 2049–2054.
Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin, 126*, 247–259.
Myrseth, K. O. R., & Fishbach, A. (2009). Self-control: A function of knowing when and how to exercise restraint. *Current Directions in Psychological Science, 18*, 247–252.
Novemsky, N., & Kronzon, S. (1999). How are base-rates used, when they are used: A comparison of additive and Bayesian models of base-rate use. *Journal of Behavioral Decision Making, 12*, 55–69.
Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review, 11*, 988–1010.
Osman, M. (2013). A case study: Dual-process theories of higher cognition—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science, 8*, 248–252.

Paxton, J. M., Unger, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgement. *Cognitive Science, 36*, 163–177.
Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014a). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition, 42*, 1–10.
Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014b). The role of analytic thinking in moral judgements and values. *Thinking & Reasoning, 20*, 188–214.
Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2013). Belief bias during reasoning among religious believers and skeptics. *Psychonomic Bulletin & Review, 20*, 806–811.
Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition, 123*, 335–346.
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. *Psychonomic Bulletin & Review, 19*, 528–534.
Pennycook, G., Trippas, D., Handley, S. J., & Thompson, V. A. (2014). Base-rates: Both neglected and intuitive. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 544–554.
Prado, J., Kaliuzhna, M., Cheylus, A., & Noveck, I. A. (2008). Overcoming perceptual features in logical reasoning: An event-related potentials study. *Neuropsychologica, 46*, 2629–2637.
Prado, J., & Noveck, I. A. (2007). Overcoming perceptual features in logical reasoning: A parametric fMRI study. *Journal of Cognitive Neuroscience, 19*, 642–657.
Rossi, S., Cassotti, M., Agogué, M., & De Neys, W. (2013). Development of substitution bias sensitivity: Are adolescents happy fools? *Proceedings of the Annual Meeting of the Cognitive Science Society, 35*, 3321–3326.
Rozyman, E. B., Landy, J. F., & Goodwin, G. P. (2014). Are good reasoners more incest-friendly? Trait cognitive reflection predicts selective moralization in a sample of American adults. *Judgment and Decision Making, 9*, 176–190.
Sá, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology, 91*, 497–510.
Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision making in the ultimatum game. *Science, 300*, 1755–1758.
Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology, 14*, 332–348.
Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in god. *Journal of Experimental Psychology: General, 141*, 423–428.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*, 3–22.
Sloman, S. A. (2002). Two systems of reasoning. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 379–398). Cambridge, MA: Cambridge University Press.
Sloman, S. A. (2014). Two systems of reasoning: An update. In J. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual process theories of the social mind*. Guilford Press.
Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4*, 108–131.
Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. Mahwah, NJ: Lawrence Erlbaum Associates.
Stanovich, K. E. (2004). *The Robot's rebellion: Finding meaning in the age of Darwin*. Chicago: The University of Chicago Press.
Stanovich, K. E. (2009b). *What intelligence tests miss: The psychology of rational thought*. London, UK: Yale University Press.
Stanovich, K. E. (2011). *Rationality and the reflective mind*. New York, NY: Oxford University Press.
Stanovich, K. E. (2009a). Is it time for a tri-process theory? Distinguishing the reflective and algorithmic mind. In J. St. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford, UK: Oxford University Press.
Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open minded thinking. *Journal of Educational Psychology, 89*, 342–357.
Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127*, 161–188.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*, 645–726.
Stanovich, K. E., & West, R. F. (2007). Natural myside 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.
Stollstorff, M., Vartanian, O., & Goel, V. (2012). Levels of conflict in reasoning modulate right lateral prefrontal cortex. *Brain Research, 1428*, 24–32.
Strack, F., & Deustch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8*, 220–247.
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.
Stupple, E. J. N., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual process account from response times and confidence ratings. *Thinking & Reasoning, 19*, 54–77.
Suskind, R. (2004). Faith, certainty and the Presidency of George W. Bush. *The New York Times*, October 17. <http://www.nytimes.com/2004/10/17/magazine/17BUSH.html>.
Svedholm-Häkkinen, A. (2015). Highly reflective reasoners show no signs of belief inhibition. *Acta Psychologica, 154*, 69–76.
Thaler, R. H., & Shefrin, H. M. (1981). An economic-theory of self-control. *Journal of Political Economy, 89*, 392–406.
Thompson, V. A. (2013). Why it matters: The implications of autonomous processes for dual-process theories—Commentary on Evans & Stanovich (2013). *Perspectives on Psychological Science, 8*, 253–256.
Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond*. Oxford: Oxford University Press.
Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning, 20*, 215–244.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, metacognition, and reason. *Cognitive Psychology, 63*, 107–140.
Thompson, V. A., Prowse Turner, J., Pennycook, G., Ball, L., Brack, H., Ophir, Y., et al (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*, 237–251.
Thompson, V. (2014). What intuitions are ... and are not. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 35–75). Burlington: Academic Press.
Toplak, M. V., 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.
Trippas, D., Verde, M. F., & Handley, S. J. (2014). Using forced choice to test belief bias in syllogistic reasoning. *Cognition, 133*, 586–600.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.
Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between Operation Span and Raven. *Intelligence, 33*, 67–81.
Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*, 104–132.
Vadeboncoeur, I., & Markovits, H. (1999). The effect of instructions and information retrieval on accepting the premises in a conditional reasoning task. *Thinking & Reasoning, 5*, 97–113.
Villejoubert, G. (2009). Are representativeness judgments automatic and rapid? The effect of time pressure on the conjunction fallacy. *Proceedings of the Annual Meeting of the Cognitive Science society, 30*, 2980–2985.
Vohs, K. D., & Heatherton, T. F. (2000). Self-regulatory failure: A resource-depletion approach. *Psychological Science, 11*, 249–254.
Wason, P. C., & Evans, J. St. B. T. (1975). Dual processes in reasoning? *Cognition, 3*, 141–154.
Whittlesea, B. W. A., Jacoby, L. J., & Girard, K. (1990). Illusions of immediate memory: Evidence of an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language, 29*, 716–732.
Whittlesea, B. W. A., & Leboe, J. P. (2003). Two fluency heuristics (and how to tell them apart). *Journal of Memory and Language, 49*, 62–79.
Wilkins, M. C. (1928). The effect of changed material on the ability to do formal syllogistic reasoning. *Archives of Psychology, 102*.
Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition, 112*, 97–104.