Acta Psychologica 187 (2018) 54–65

Contents lists available at ScienceDirect

Acta Psychologica

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

To detect or not to detect: A replication and extension of the three-stage model☆

Alexander B. Swan[a,c,*], Dustin P. Calvillo[b], Russell Revlin[a]

[a] Department of Psychological and Brain Sciences, University of California, Santa Barbara, United States
[b] Department of Psychology, California State University San Marcos, United States
[c] Social Science and Business Division, Eureka College, United States

ARTICLE INFO

Keywords:
Dual process theory
Conflict detection
Base-rate neglect
Conditional reasoning

ABSTRACT

When faced with a decision, people generally show a bias toward heuristic processing, even if it leads to the incorrect decision, such as in the base-rate neglect task. The crucial question is whether people know that they are biased. Recently, the three-stage model (Pennycook, Fugelsang, & Koehler, 2015) suggested that detecting this bias (*conflict detection*) is imperfect and a consistent source of bias because some people do not recognize that they are making biased decisions. In Experiment 1, participants completed a base-rate neglect task as replication of Pennycook et al. (2015). In Experiment 2, a conditional reasoning task was added as an extension to test the boundary conditions of the model. Results in Experiment 1 indicated that detection failures were a significant source of bias. However, results in Experiment 2 on the conditional reasoning task indicated that the three-stage model may be incompatible with a complex task such as conditional reasoning, an issue explored in detail in the General discussion.

1. Introduction

Humans can be poor thinkers, especially when a reasoning problem or situation requires logic or probability computations. They often rely on beliefs and other heuristics when making decisions rather than considering logic or probability. For example, in a 2016 episode of *Last Week Tonight with John Oliver*, Mr. Oliver described in excruciating detail the central, biased problem with science communication and consumption in our society: too much bias. Perhaps the summative point came from a clip of TV weatherman Al Roker, who suggested that viewers should merely find a scientific study that confirms their beliefs and use it as evidence to justify their behaviors (Oliver & Pennolino, 2016). Of course, this is usually a bad idea and shows the tendency to rely on one's beliefs rather than critically evaluating evidence.

So why do people fall victim to biased thinking? Perhaps the answer lies in the idea that there may be two types of information processing, and much of the susceptibility toward bias is due to the issues associated with one or both of those processing routes. Though they are varied in their application and description, the main distinction in *dual process* theories of thinking is that there is one fast, heuristic, and automatic processing route (Type 1, or T1) and one slow, deliberate, and analytic processing route (Type 2, or T2) (e.g., De Neys, 2012; Evans, 2003, 2007; Evans & Stanovich, 2013; Frankish & Evans, 2009; Stanovich, 2009). Within the realms of problem solving, decision-making, and reasoning, a given problem may contain cues or information that may engender *conflict* between these two routes of processing. In other words, are the two routes signaling two different answers from the initial understanding of the problem? One of the central arguments within the individual differences literature in dual process thinking is whether and under what circumstances people *detect* that conflict, and how might they *resolve* it (De Neys, 2012; Pennycook, Fugelsang, & Koehler, 2015).

Recently, Pennycook et al. (2015) described a three-stage model of conflict detection. The authors reframe the conflict as not merely a T1–T2 conflict, but more appropriately a conflict of generated responses in T1 (initial responses that reflect distinct decisions). They also argued that T2 processing, as an imperfect analytic processing system, should have measurable differences in quality depending on the judgment or reasoning task.

To illustrate a better picture of the model and what is meant by quality, let's look at the stages. The first stage of the three-stage model is purely T1 processing, which generates an initial response upon an initial reading of problem or situation. A person can generate multiple initial responses. The decision time in this stage is characterized in

☆ Portions of this manuscript appear in the doctoral dissertation of Alexander B. Swan. The authors would like to thank Mary Hegarty, Richard E. Mayer, Eric R. A. N. Smith, Wim De Neys, Gordon Pennycook, and two anonymous reviewers for their thoughtful questions, comments, and suggestions on earlier versions of this manuscript.

* Corresponding author at: Eureka College, 300 E College Ave, Eureka, IL 61530, United States.
E-mail address: aswan@eureka.edu (A.B. Swan).

https://doi.org/10.1016/j.actpsy.2018.05.003
Received 12 October 2017; Received in revised form 3 May 2018; Accepted 9 May 2018

milliseconds. Stage 2 is the conflict monitoring/detection stage. If the problem or situation engenders a conflict in outputs, then stage 3 is entered and T2 processing is engaged. This monitoring/detection process is also characterized in milliseconds. If there is no conflict, then stage 3 is merely the output of the first initial response.

In all cases, stage 3 is T2 engagement and reflects the response output. Pennycook et al. (2015) make two qualitative distinctions in this final stage: *rationalization* and *cognitive decoupling*. If the stage 1 (initial) response is ultimately chosen, then the person is said to have rationalized: reasons for justification of this response are made, even if it is not the normative response. Alternatively, cognitive decoupling refers to the decision on a problem where an initial response is suppressed for the alternative. Taken together, this stage is reflected by slower responding, with final decisions taking seconds of processing, rather than milliseconds as in the first two stages. Ultimately, they argued that bias on judgment tasks, including base-rate neglect, can occur at any stage, including the monitoring in stage 2 or after detection in stage 3.

The primary task utilized in Pennycook et al. (2015) was the *base-rate neglect* task. Sometimes referred to as the "lawyer-engineer" problem, this task uses the tendency of people to ignore numerical information (base-rates) when making judgments about group membership, due to a reliance on the representativeness heuristic (Tversky & Kahneman, 1974). For example, imagine personality descriptions were available for 100 people, 70 lawyers and 30 engineers. One individual is "randomly selected" from the sample that has a description that sounds like a stereotypical engineer. A participant then determines to which group this individual most likely belongs. Base-rate neglect arises if the participant responds with the answer "engineer", even though the lawyer group had larger base-rate in the sample. This is due to incongruity of the base-rates and the description of the person. If the base-rates were switched, and there were 70 engineers and 30 lawyers, but the description of the individual remained the same, participants tend to choose the larger group category quickly, as the stereotype and base-rate responses align.

De Neys and Glumicic (2008) proposed a dual process explanation of the base-rate neglect task. Just as Tversky and Kahneman (1974) found that the answer cued by a stereotypical description would be chosen more often than the base-rate answer on problems where the base-rates and descriptions are incongruent, De Neys and Glumicic argued that this occurs because of the tendency of the participants to rely upon T1 processing and not entertain probabilities, which may require T2 processing. Second, they argued that a slower response time would be indicative of the activation of analytic thinking and T2 deliberation. Moreover, the longest response times would reflect the idea that the incongruence/conflict between the base-rate answer and the more salient stereotype answer had been detected and the proper inhibition of that response was necessary to arrive at the base-rate answer. Third, this conflict detection would be measureable by asking participants to recall the base-rate values after the main task, wherein incongruent problems would show better recall performance. This would be an implicit measure of base-rate encoding; a verbal protocol procedure was also used to measure explicit encoding or use of the base-rates. De Neys and Glumicic found support for their predictions (but no support for the explicit recognition of base-rates during the verbal protocol portion), and concluded that a conflict detection system operates constantly and operation is efficient and routine. They also concluded that, while this conflict detection system works well, and describes the modal, or average, reasoner, it's not a perfectly reliable system and individual differences cannot fit into the predictions of their model (De Neys & Glumicic, 2008; see De Neys, 2012, 2014, for further developments).

These conclusions have led to an increased focus on the conflict detection mechanism's role within dual process theories. Other researchers have investigated base-rate neglect and have supported and advanced De Neys and Glumicic's (2008) initial conclusions. Pennycook, Fugelsang, and Koehler (2012) replicated De Neys and Glumicic (2008) with the extreme base-rate values used (e.g., 997 vs. 3), but could not replicate the findings when using the original base-rates (described above, 70 vs. 30). However, they did conclude that conflict detection as a monitoring system does operate at some level within T1, even before T2 thinking is engaged. Pennycook and Thompson (2012) further argued that a person's utilization of base-rates is not within the domain of T2, but under the purview of T1. These results showed that the use of the base-rates within a problem is as effortless as incorporating the stereotypic information, which has always been assumed to be a T1 process (see also Pennycook, Trippas, Handley, & Thompson, 2014).

1.1. Present experiments

The following two experiments represent a replication of the initial study that tested the three-stage model (Experiment 1), as well as an extension (Experiment 2).

In Experiment 1, participants completed a base-rate neglect task, combining methodologies of Pennycook et al. (2015) and De Neys and Glumicic (2008). The two main predictions of the model were tested. First, reasoners might be biased because of *monitoring/detection failures*. Monitoring/detection failures occur when reasoners have the necessary formal knowledge, but are unaware that the problem demands it (De Neys & Bonnefon, 2013). The three-stage model claims that poor performers on the task cannot reliably detect conflicts. Therefore, it predicts a positive relationship between the amount of base-rate responses and response time differences between initial stereotype responses and alternative base-rate responses (*conflict detection hypothesis*). This latter measure is defined as the *conflict detection index*. Second, the effort of T2 processing was tested: the cognitive decoupling hypothesis is described as the response time difference between base-rate responses on conflict problems from the nonconflict response time baseline, which represents the progression of successful conflict detection to the alternative (base-rate) response on conflict problems (*the decoupling index*). The relationship between base-rate responses on conflict problems and this response time difference should be negative, as participants who choose the base-rate less often must decouple to arrive at the base-rate answer. This response time difference should decrease as base-rate decisions increase.

In Experiment 2, the three-stage model predictions (the conflict detection and cognitive decoupling hypotheses) were extended to a conditional reasoning task (Thompson, 1994). The three-stage model was not intended to be only a base-rate neglect model (Pennycook et al., 2015), and thus boundary condition testing (De Neys, 2014) was warranted here. Dual process conflict detection mechanism predictions have been tested on various tasks, such as the conjunction fallacy (De Neys, Cromheeke, & Osman, 2011), syllogisms (De Neys, Moyens, & Vansteenwegen, 2010), ratio bias (Mevel et al., 2015), the bat and ball problem (De Neys, Rossi, & Houdé, 2013; Johnson, Tubau, & De Neys, 2016), number conservation (De Neys, Lubin, & Houdé, 2014), and math reversal errors (Lubin, Houdé, & De Neys, 2015). Additionally, the three-stage model has been supported in recent studies (e.g., Bago & De Neys, 2017; Frey, Johnson, & De Neys, 2017; Newman, Gibb, & Thompson, 2017). Frey et al.'s findings supported the conclusions that conflict detection is an individual difference. Bago and De Neys directly supports the three-stage models predictions and outcomes, while Newman et al. supports the notion that conflict arises from competing T1 responses. Newman et al. corroborated this idea on both base-rate and conditional reasoning tasks, but it is worth noting that the conditional task did not directly align with the three-stage model predictions, which the focus in Experiment 2. The addition of conditional reasoning allowed for a contrast of logic principles and believability, another situation of heuristic knowledge competing with the normative principles of logic. A comparison of conflict detection was examined within and between both tasks (base-rate and conditional reasoning) of this

experiment.

2. Experiment 1

Experiment 1 was conducted to test the major behavioral predictions of the three-stage model. It states that conflict detection is a fast, T1 process, but that it is bias-prone. The proportion of base-rate responses and response times (RT) are crucial to describing the conflict detection and monitoring claims of this model.

First, participants will choose the stereotype response more often in conflict problems (where the stereotype answer is different from the base-rate answer) than in nonconflict problems (where the stereotype and base-rate answers are the same). This prediction reflects the conventional and robust effect of the base-rate neglect task.

Second, if there is a conflict detection process that people use, then stereotype responses on conflict problems will be slower than on base-rate nonconflict problems indicating general conflict detection and support for the model (*conflict detection hypothesis*). This relationship would be positive, suggesting that stereotype-biased participants do not detect conflict as easily as participants with greater base-rate responses.

Third, the three-stage model makes a distinction with base-rate responses on conflict problems: the longest RTs reflect cognitive decoupling, or choosing the alternative (base-rate) response (Pennycook et al., 2015), and this is reflected by the RT difference between average base-rate nonconflict responses and base-rate conflict responses. Cognitive decoupling responses should take longer than stereotype conflict responses, but only if conflict is detected. Per the three-stage model, this latter process is described as rationalization. Because rationalization relates to a stereotype answer in this paradigm, greater RT but low base-rate responses would indicate this; base-rate answers on conflict problems should show a decrease in RT as the number of base-rate responses increase (*cognitive decoupling hypothesis*).

In the case of the conflict monitoring claim, base-rate nonconflict RTs are used as a baseline measure to create RT difference scores. The reason RTs on nonconflict problems represent a baseline is because the problems generally have one answer cued by both the base-rates and the stereotypic information in the problem. Historically, base-rate responses are generally at a ceiling (e.g., De Neys & Glumicic, 2008; Pennycook et al., 2012; Pennycook et al., 2015). Since a single answer is cued by both sources of information in the problems, RTs are assumed to be extremely quick beyond initial reading time, acting as starting point for measuring additional processing that is hypothesized to reflect conflict detection and T2 engagement.

2.1. Method

2.1.1. Participants

Ninety-three psychology undergraduate students (71% female; M_{age} = 18.85 years, SD = 1.44) participated in this study for partial course credit. A sensitivity analysis for this sample size was conducted with the following parameters: one-tailed $\alpha = 0.05$ and power 0.80; the test revealed that with the sample size of 93, we were well powered to examine effect sizes of Cohen's $d = 0.26$ or higher for the experimental manipulations and $r = 0.25$ or higher for the correlational analyses. Effects observed are discussed within this sensitivity.

2.1.2. Design and materials

2.1.2.1. Base-rate task. The materials used for this study were adapted from De Neys, Vartanian, and Goel (2008). A typical problem appeared like this:

> In a study 1000 people were tested. Among the participants there were 5 men and 995 women. Jo is a randomly chosen participant of this study.
>
> Jo is 23 years old and is finishing a degree in engineering. On Friday nights, Jo likes to go out cruising with friends while listening to loud music and drinking beer.
>
> What is most likely?
> a. Jo is a man
> b. Jo is a woman

This is an example of a conflict problem (base-rate information and diagnostic information are incongruent). The other type of problem that participants answered was nonconflict problems, where the base-rate information and diagnostic information about the randomly selected individual were congruent with one another (swap the numbers for men and women in the above problem). Participants answered 50 total base-rate problems, with 25 conflict problems and 25 nonconflict problems. The neutral condition from De Neys and Glumicic (2008), in which problems do not contain stereotypic information, was not included to maximize the contrast in problems type. Stereotypes used varied in content: age, gender, race, job-related groups, and stereotypical human characteristics.

In addition to the congruency of the base-rate information and the diagnostic information, there were three expressions of extreme base-rates (997 to 3, 996 to 4, and 995 to 5). This varies the presentation of the problems, as well as to force reading of the base-rate information. Previous research using this methodology has shown that the extreme base-rates are needed for contrast between the conflict and nonconflict problems (De Neys & Glumicic, 2008; Pennycook et al., 2012) vs. the traditional 70/30 split used in Tversky and Kahneman (1974).

The task began with an overview of the fake survey that participants believed was the basis for the task. For each problem, they were told that a random sample of 1000 respondents from the survey was selected. Below is the overview:

> In this stage, you will encounter 50 problems regarding a recent comprehensive survey that was conducted in the country. Various pieces of information were gathered from thousands of individuals. Each problem will have a random subset of 1000 responses.
> For each problem, you will be given a brief description of a randomly selected individual from the subset. **Based on the information provided, it is your task to decide to which group the individual belongs.**

2.1.2.2. Individual difference measures. In addition to the main base-rate decision-making task, participants completed three individual differences measures. These measures included the Cognitive Reflection Test (Frederick, 2005), the Need for Cognition scale (NFC; Cacioppo, Petty, & Kao, 1984) and the Actively Open-minded Thinking scale (AOT; Stanovich & West, 1997), as well as utilizing SAT scores as a cognitive ability measure (e.g., Stanovich & West, 2000). Analyses and discussion of these tasks can be found in the Supplemental Analyses document.

2.1.3. Procedure

Participants completed each of the measures at a computer station. Each student was introduced to the study and told that there were four stages to the entire session. Participants first solved the three problems of the CRT. After this, they answered the 50 base-rate problems. The order of these problems was fully randomized. Additionally, the base-rate answer was randomized, and it was either presented as the first option or second option. Once the participants finished with those problems, they were given an opportunity to rest. After the short rest period, instructions for the NFC appeared, describing the questionnaire and each of the corresponding scale values. The question order was randomized, and the scale appeared below each question. Upon completion of the NFC, the AOT was presented. Instructions and description of the scale preceded the questions. Question order was randomized and the scale appeared below each question. Finally, participants

Table 1
Mean (SD) proportion of base-rate responses and response times (s) in Experiment 1.

	Prop. base-rate responses	Response time
Problem type	*M* (*SD*)	*M* (*SD*)
Nonconflict	0.90 (0.08)	11.6 (3.04)
Conflict	0.48 (0.30)	
Stereotype responses		14.3 (4.91)
Base-rate responses		14.5 (5.21)

entered their most recent SAT score (out of 2400) and some demographic information. Participants were then debriefed, thanked, and dismissed.

2.2. Results

Behavioral analyses were conducted to test the model predictions. In all analyses described below, if a participant did not contribute a value to all cells associated with the specific statistical test (i.e., did not make any errors on the task), the participant was excluded. Specific exclusions are noted for the corresponding tests.

First, it was predicted nonconflict problems would garner higher base-rate responses than conflict problems. This prediction was supported. As shown in Table 1, participants chose the base-rate answer on nonconflict problems more often than on the conflict problems, $t(92) = 15.47$, $p < .001$, Cohen's $d = 1.47$. The variance was small in this set of 25 nonconflict problems; three problems could be considered outliers for base-rate responses values below 2.5 *SD*s below the overall mean of set.[1] The distribution of problem means for conflict problems was larger, but no problems fell outside of the outlier range. For many of the conflict problems, the stereotype decision was made significantly more often than chance performance. Though the proportion of base-rate responses for conflict problems was approximately 40%, higher than historically measured (e.g., De Neys & Glumicic, 2008), this was likely due to an overall marginal practice effect, as participants evaluated many more problems than typically presented in previous studies. However, this methodological practice is not without precedent (De Neys et al., 2008) and time series analyses did not indicate a significant increase from initial problems to later ones to suggest this to be an issue.[2]

It was predicted that there would be an increase in RTs[3] for stereotype conflict responses above a baseline base-rate nonconflict RT (*conflict detection hypothesis*). This simple increase marks the presence of overall conflict detection within the group. As Table 1 reveals, participants took significantly longer to answer conflict problems with the stereotype response than nonconflict problems overall, $t(84) = 5.44$, $p < .001$, $d = 0.62$.[4] However, this overall difference needs further explanation by incorporating the proportion of base-rate responses.

In contrast, the converse piece of the three-stage model was also tested, which suggests that participants with numerous stereotype responses on conflict problems are less likely to show conflict detection within their responses. For each participant, mean base-rate RT on nonconflict problems was subtracted from RT on stereotype conflict responses. This RT difference value represents the *conflict detection index*. Fig. 1 illustrates this index along with the other index described below. There was a large, positive correlation between conflict base-rate responses and the conflict detection index, $r(83) = 0.69$, $p < .001$, $R^2 = 0.48$, indicating that as the number of base-rate responses on the conflict problems increased, so did the RT difference between stereotype responses on those same conflict problems and nonconflict problems. People who selected the base-rates more often on the conflict problems had greater RT differences than people who biased by the stereotype on conflict problems. Moreover, there was a large negative intercept, $b = -2.02$ s, $t(84) = -3.05$, $p = .003$. This outcome supports the three-stage model and replicates the findings of Pennycook et al. (2015). Some individuals did not seem to show conflict detection within their RTs and conflict detection does appear to be an individual difference in this group of participants.

Finally, the behavioral prediction of *cognitive decoupling* was tested. The cognitive decoupling index was computed by subtracting nonconflict base-rate RTs from conflict base-rate RTs. Positive values reflect greater time processing the base-rate answer on a conflict problem than when information-cues in the problem are congruent. This reflection is cognitive decoupling, especially indicative on poor performers who give a periodic correct answer (less than chance). The relationship between this index and conflict problem accuracy should be negative to support the three-stage model prediction. Though the relationship was negative, it was not significant, $r(83) = -0.16$, $p = .15$, $R^2 = 0.03$. The intercept was significantly greater than zero, $b = 4.04$ s, $t(91) = 6.05$, $p < .001$, suggesting that decoupling required several seconds to reverse a stereotype pattern of responding for individuals who tended to respond in this way. See Fig. 1 for the pattern of responses on this index.

2.3. Discussion

The results of Experiment 1 indicated mixed support for the general predictions proffered by the three-stage model of conflict detection. Behaviorally, most participants neglected base-rates on conflict problems. Stereotype conflict judgments took longer than any nonconflict judgments; this difference represents conflict detection, at least implicitly—but this did not occur for all participants. This finding provides support for the three-stage model and replicates this finding specifically, which claims that conflict detection is subject to individual differences (Pennycook et al., 2015).

With respect to the decoupling hypothesis, support was mixed. There was muddied evidence for decoupling on the decoupling index. The direction of the relationship with conflict base-rate RT and nonconflict base-rate RT was in the predicted direction, but the relationship was not significant. This suggests that participants who chose the stereotype answer more frequently may have decoupled a handful of times. This does support the idea that decoupling, in the three-stage model, does indeed take longer than rationalization—though Pennycook et al. (2015) do not make any strong claims as to why this might be the case, as the central predictions of the model are concerned with the imperfect efficiency of detection. It is worth noting that the decoupling index was used primarily as a dissociative measure between base-rate responses and RT correlations, rather than a direct prediction/test of T2 responses. Regardless, the trend here continues to support the time distinction between the two aspects of stage 3/T2 thinking.

The base-rate neglect task is a social judgment task. Participants make inferences about people, not about objects or pretend situations. It is not necessarily a reasoning task per se and probabilistic judgments are distinct from logical decisions (thought they may be of the same "normative" ilk). For the three-stage model to be considered as a strong model for the explanation of conflicting initial responses on all kinds of thinking problems, it needs to explain and inform the observations on other judgment and reasoning tasks. Experiment 2 incorporated a conditional reasoning task to broaden the scope of the investigation and extend the three-stage model beyond the base-rate neglect task.

[1] All analyses were conducted without these 3 problems, but no meaningful changes to base-rate responses were observed.

[2] There was a positive trend of base-rate responses on conflict problems, but this trend was not significant over all problems ($r = 0.26$, $p = .06$).

[3] Response times (RTs) were analyzed as both raw and as $\log_{10}$ transformed. No outliers were detected; for the sake of simpler interpretation, all RTs are presented in seconds and milliseconds.

[4] Eight participants were excluded from this analysis for not contributing any stereotype conflict responses.

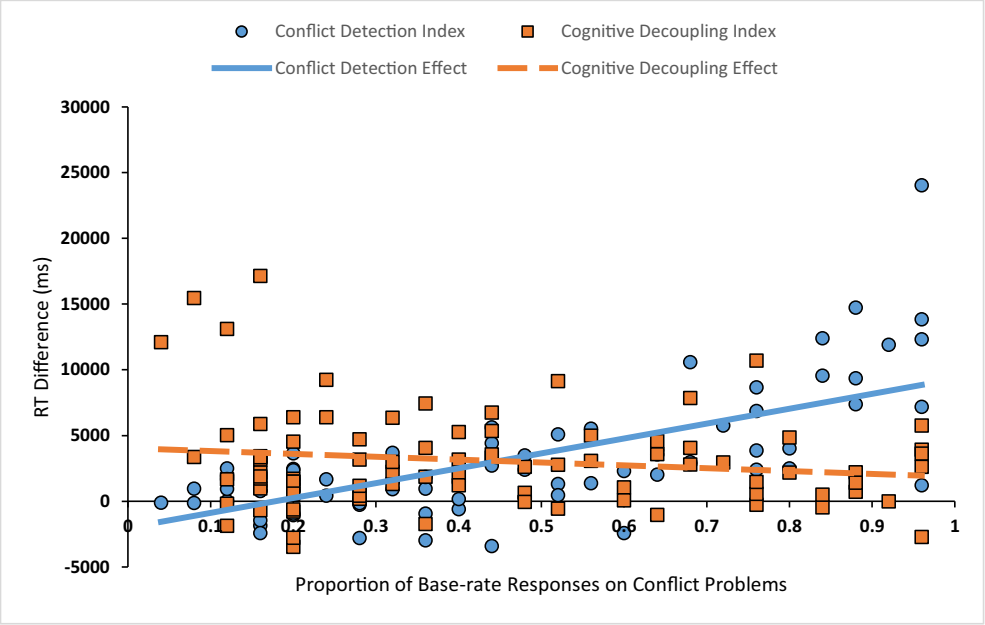

Fig. 1. Scatterplot of mean RT time differences and the proportion of base-rate responses on conflict problems in Experiment 1. The conflict detection index refers to the RT difference between stereotype conflict responses and base-rate nonconflict responses. The decoupling index refers to the RT difference between base-rate conflict responses and base-rate nonconflict responses. Each unit represents one participant (i.e., one circle and square per participant). Lines show regressions of proportion of base-rate responses on RT difference scores.

3. Experiment 2

The behavioral results of Experiment 1 showed mixed support for the three-stage model, but replicated the conflict detection results observed in Pennycook et al. (2015). The goals of Experiment 2 were as follows: (1) To test the expansion of the model by including an additional and qualitatively different task within the reasoning realm. The three-stage model was first tested on the base-rate neglect task; it is a reasonable undertaking to test the model's predictions on a different task, such as conditional reasoning (though the authors of that work do discuss what patterns might appear, Pennycook et al., 2015). Results from recent studies are also consistent with the three-stage model (e.g., Bago & De Neys, 2017; Frey et al., 2017; Newman et al., 2017). (2) To compare behavioral conflict monitoring on the conflict detection index with an additional, explicit measure of monitoring/detection: confidence in one's answer (the three-stage model is vague on metacognitive processes; see Pennycook, 2017, for further discussion).

The basic measure of confidence is related to the idea proffered by Thompson and colleagues termed Feeling of Rightness (FOR; Thompson, 2009; Thompson, Prowse Turner, & Pennycook, 2011; Thompson et al., 2013). They argued that this FOR signals T2 engagement. When confidence is low in an initial response to a problem or situation, deliberation may be required, and T2 is engaged. Conversely, if confidence is high in an initial response, T2 will likely not be engaged. This measure is inversely related to RTs, as low confidence would signal T2 engagement, thereby slowing responses; for high confidence, T1 responses will remain quick. FOR is measured in a two-response paradigm, which was not the case in Experiment 2. Thus, confidence is merely an additional, indirect measure of conflict monitoring/detection (e.g., Frey et al., 2017).

A similar methodology to Experiment 1 was used in the present experiment with a few exceptions. The same base-rate problems from the first experiment were used; a conditional reasoning task was added, utilizing some of the methodology of Thompson et al. (2011); the AOT as an individual difference measure was dropped due to its length and its high correlation with the NFC scale; and participants rated their confidence for each base-rate and conditional reasoning problem.

It was predicted that overall base-rate task performance would replicate Experiment 1. Participants will have lower base-rate responses on conflict problems than nonconflict problems, while also responding slower on those conflict problems than nonconflict problems. If the three-stage model is broad enough to encompass higher order judgments and reasoning within the dual process framework, then the base-rate pattern of results should replicate on the conditional reasoning problems. Confidence will be inversely related to response time, whereby low confidence will likely engage T2 thinking, increasing processing times and high confidence will not engage T2, decreasing processing times.

3.1. Method

3.1.1. Participants

One hundred twelve undergraduates initially participated in this experiment for partial course credit. Ten participants were dropped from all analyses due to incompletion of all experimental tasks. Thus, 102 participants (62% female; M_{age} = 18.92, SD = 1.12) were included in overall data analysis. As with Experiment 1, if participants did not contribute to the all cells in each statistical test (i.e., no stereotype or incorrect validity answers), they were excluded. Those sub-*ns* are noted for each test.

A sensitivity analysis for this sample size was conducted with the following parameters: one-tailed $\alpha = 0.05$ and power 0.80; the test revealed that with the sample size of 93, we were well powered to examine effect sizes of 0.24 or higher for the experimental manipulations and $r = 0.23$ or higher for the correlational analyses. Effects observed are discussed within this sensitivity.

3.1.2. Design and materials

3.1.2.1. Base-rate task. This was the same as Experiment 1.

3.1.2.2. Conditional reasoning task. The conditional reasoning task was drawn from Thompson et al. (2011). The conditional reasoning task asked participants to complete an inference (drawing a conclusion) when the initial statement is presented in the form of "*if p, then q*". Four inferences can be made from this single form: Modus Ponens (MP), Modus Tollens (MT), Affirming the Consequent (AC), and Denying the Antecedent (DA). The first two inferences are logically valid (i.e., the conclusion follows necessarily from the premises), and the second two inferences are logically invalid. An example is as follows:

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

Table 2
Mean (SD) proportion of base-rate responses and response times (s) on the base-rate neglect and conditional reasoning tasks in Experiment 2.

	Base-rate neglect task		Conditional reasoning task	
	Prop. of base-rate responses	Response time	Prop. of base-rate responses	Response time
Problem type	n = 102	n = 93	n = 102	n = 102
Nonconflict	0.90 (0.11)	13.1 (3.47)	0.77 (0.12)	8.1 (2.37)
Conflict	0.45 (0.27)		0.40 (0.15)	
Stereotype resp./ incorrect		14.4 (4.00)		7.9 (2.45)
Base-rate resp./ correct		16.4 (5.07)		8.8 (3.46)

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

There were 64 total problems, with 16 problems of each inference. To create nonconflict and conflict problems, believability was manipulated. Half of the problems were believable, where p was a sufficient condition to bring about q in valid inferences or when p was a necessary condition to bring about q in invalid inferences; the other half of the problems were unbelievable, where p was not sufficient for q for valid inferences, or it was not at all necessary in the case of invalid inferences (Thompson, 1994). Thus, there were eight problems in each set of 16 that were believable and there were eight that were unbelievable. The problems either represented a causal conditional or a definitional conditional. All problems used for this task were originally developed in Thompson (1994).

The instructions for the conditional reasoning task directed the participants to think logically about each problem, and to determine whether the conclusion *logically followed* from the set of two premises. For each conditional reasoning problem, the answer prompt was worded to mirror the base-rate task. For example, it appeared after the presentation of the two premises like this (from MP above): "*Is the conclusion likely to follow from the premises?*" This question was followed by a simple YES or NO response. A correct answer for valid conditionals was a YES response; for invalid conditionals, it was a NO response. The results were coded as such for the dependent variable of accuracy the conditional reasoning task only.

3.1.2.3. Confidence. Confidence ratings were added in this experiment as an additional measure of conflict monitoring/detection, based on previous studies of FOR (e.g., Thompson et al., 2011). Participants responded to the question, "At the time I provided my answer, I felt ___." They responded on Likert scale ranging from 1 (*guessing*) to 7 (*certainty*), with the midpoint of the scale labeled *fairly certain*. This rating was given after each problem of both tasks and recorded a total of 114 times.

3.1.2.4. Thinking disposition and cognitive ability measures. Similar to Experiment 1, this experiment used the CRT, NFC, and the SAT score. The analyses and discussion of these measures is available in the Supplemental materials.

3.1.3. Procedure

Participants completed the each of the tasks at a computer station. Each student was introduced to the study and told that there were four stages to the experimental session. Participants first solved the three problems of the CRT. Participants then answered the 50 base-rate problems. The order of these problems was fully randomized. Additionally, the correct answer was randomized, and it was either presented as the first option or second option. Once participants finished with those problems, they were given an opportunity to rest for 30 s before answering the 64 conditional reasoning problems. The order of these questions was fully randomized. After participants finished these problems, they were given an additional opportunity to rest for 30 s before completing the NFC and providing their most recent SAT score (out of 2400) and demographic information. Participants were debriefed, thanked, and dismissed. The entire experimental session lasted approximately 45–60 min.

3.2. Results: Base-rate neglect task

The base-rate task was analyzed the same as Experiment 1. The prediction was a replication of the robust representativeness heuristic on conflict problems. Table 2 shows the means (*SD*s) for responses and reveals that this prediction was supported: nonconflict problems had reliably higher base-rate selections than conflict problems, $t(101) = 19.20$, $p < .001$, $d = 1.86$.

3.2.1. Response time

Table 2 also shows that the prediction that conflict problems would be slower than nonconflict problems was supported, $t(101) = 8.88$, $p < .001$, $d = 0.41$. There was a significant main effect on RT across responses, $F(2, 184) = 49.53$, $p < .001$, $\eta_p^2 = 0.35$.[5] Moreover, the overall RTs for each in planned pair comparisons were reliably different (Bonferroni correction, all pairs, $p < .001$). There was a large difference between conflict base-rate responses and conflict stereotype responses, with the former leading to two additional seconds of processing, on average. This large two-second processing gap supports the three-stage model's claims that T2 processing functions at seconds-level processing.

3.2.1.1. Conflict detection index. Fig. 2 illustrates the positive relationship between the conflict detection index (conflict stereotype RT – nonconflict base-rate RT) and the proportion of base-rate responses on conflict problems. At the group-level, as stated above, the two measures in the index had distinct time stamps, $t(92) = 4.62$, $p < .001$, $d = 0.33$. To replicate support for the three-stage model, a positive correlation for that relationship was required. A positive correlation was observed, $r(91) = 0.37$, $p < .001$, $R^2 = 0.14$. Additionally, the intercept of this relationship was not significantly different from zero, which represents the lack of conflict detection or a mechanistic monitoring failure, $b = -266$ ms, $t(91) = -0.52$, $p = .60$, for a predicted individual who did not get a single problem correct on conflict problems. This relationship replicated the effect observed in Experiment 1.

3.2.1.2. Cognitive decoupling index. The cognitive decoupling index (conflict base-rate RT – nonconflict base-rate RT) should show a negative relationship between this RT difference and base-rate responses on conflict problems to support the three-stage model. Fig. 2 shows that there was a nonsignificant negative relationship between the two measures, $r(91) = -0.06$, $p = .59$, $R^2 = 0.003$. The intercept was significantly different from zero, $b = 3.80$ s, $t(91) = 5.64$, $p < .001$, suggesting that base-rate RTs are usually marked by positive differences with nonconflict RTs. Again, this finding does not necessarily support the three-stage model, though the index is in the

[5] This test included 93 participants. Nine participants were excluded because they did not make any errors on conflict problems.

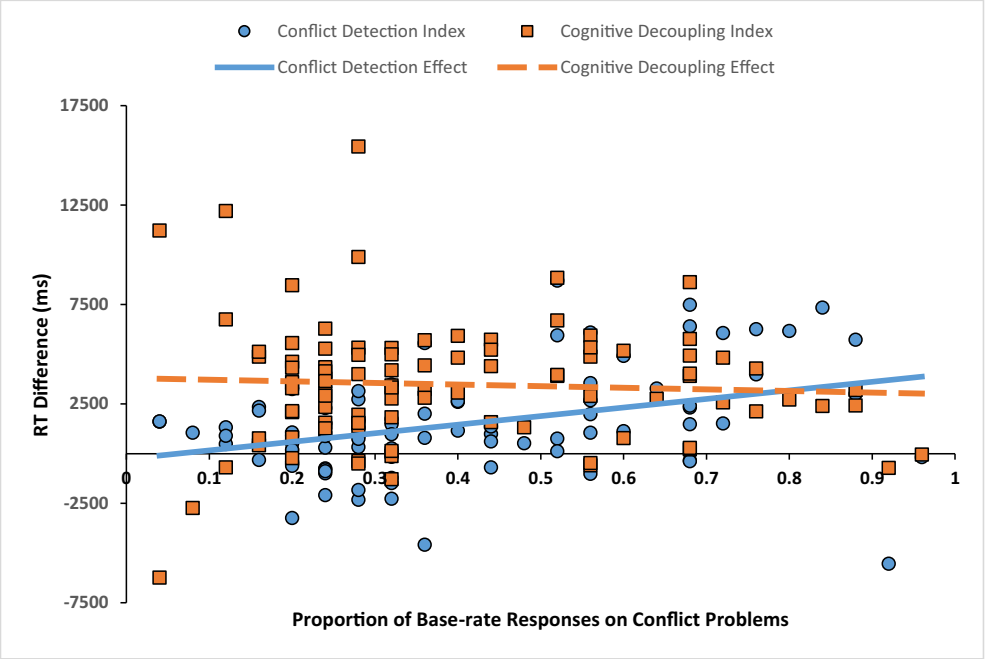

Fig. 2. Scatterplot of mean RT time differences and the proportion of base-rate responses on conflict problems the base-rate neglect task in Experiment 2. The conflict detection index refers to the RT difference between conflict stereotype responses and overall nonconflict responses. The decoupling index refers to the RT difference between conflict base-rate responses and overall nonconflict responses. Each unit represents one participant (i.e., one circle and square per participant). Lines show regressions of proportion of base-rate responses on RT difference scores.

appropriate direction. It is difficult to support the strong decoupling claims and findings in Pennycook et al. (2015).

3.2.2. Confidence ratings

The three-stage model aligns with the claim that if a person has fluency of answer generation, then the answer comes to mind fast (T1) and should yield high confidence. Alternatively, if T1 generates multiple responses initially, then monitoring should produce lower confidence if the conflict is detected. These claims were tested in two ways: 1) correlations were computed between confidence ratings and behavioral conflict monitoring (conflict detection index) and 2) with a similar index computation of confidence ratings that parallels the RT conflict detection index and comparing that with proportion of base-rate responses on conflict problems. The former test should maintain a negative relationship: as confidence increases, then the conflict detection index value should decrease. If reasoners do not detect a conflict, their conflict detection index value will be low, but there will likely be high confidence. The latter test should yield a negative correlation/regression effect: if reasoners do not detect a conflict, their confidence will be high; if a conflict is detected, proportion of base-rates will likely increase (as indicated by the behavioral effects), and decreasing confidence ratings overall.

The first test of confidence ratings yielded marginal, negative correlations (shown in Table 3) between confidence and the detection index. Nonconflict problems had a stronger correlation overall, $r(91) = -0.17$, $p = .055$, $R^2 = 0.03$ one-tailed (marginal, correct direction). A more conclusive relationship would be the relationship on conflict problems, but this relationship was not significant, $r(91) = -0.15$, $p = .08$, $R^2 = 0.02$, one-tailed (however, correct direction). This is mixed converging evidence, but these correlation analyses are potentially underpowered as indicated by the sensitivity analyses.

A confidence rating conflict detection index was computed following a similar subtraction method for RT values to complete the second test. Base-rate confidence ratings on nonconflict problems were subtracted from stereotype confidence ratings on conflict problems. A bivariate regression analysis revealed that as proportion of base-rate responses on conflict problems increase, the confidence rating conflict index decreased, $r(91) = -0.51$, $p < .001$, $R^2 = 0.26$. Thus, as participants made more base-rate responses on conflict problems, their confidence ratings began to dip from how confident they were on problems with no conflict. This finding, and the comparison of ratings to the RT conflict detection, support the overall role of confidence in one's answers in generating conflict and some individuals' ability to resolve the conflict. Furthermore, these findings additionally support the notion of the three-stage model that conflict detection is an individual difference, and some individuals might be more confident of

Table 3
Correlations between task responses and confidence measures in Experiment 2.

Measure	1	2	3	4	5	6	7	8	*M*	*SD*
1. BR nonconflict base-rate responses	–								0.90	0.11
2. BR conflict base-rate responses	.52***	–							0.45	0.27
3. BR nonconflict confidence	.20*	.31**	–						4.69	1.26
4. BR conflict confidence	−0.02	.19†	.89***	–					4.23	1.26
5. CR nonconflict accuracy	0.10	0.05	0.02	−0.11	–				0.77	0.12
6. CR conflict accuracy	0.07	−0.05	0.14	0.17	−.33**	–			0.40	0.15
7. CR nonconflict confidence	0.07	0.12	.64***	.59***	.19†	0.06	–		5.29	1.18
8. CR conflict confidence	0.12	0.14	.69***	.62***	.16	0.08	.97***	–	5.24	1.18

Note. BR = base-rate neglect task; CR = conditional reasoning task.
† $p < .10$.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

their stereotypic decisions when there is conflict in initial responses.

Additional confidence ratings tests are included in the Supplemental materials.

3.3. Results: Conditional reasoning task

Analysis of the conditional reasoning proceeded identically to the base-rate neglect task. An effort is made in this section to discuss results of the task and dependent measures within this task, as well as a comparison to the results of the base-rate neglect to track if patterns of behavior and cognitive processes extend to this traditional reasoning task.

After combining the problem types into nonconflict and conflict to match the base-rate neglect task, similar analyses were conducted.[6] Table 2 displays the means (*SD*s) and reveals that was an accuracy effect, as participants were generally more accurate on nonconflict problems than conflict problems, $t(101) = 17.25$, $p < .001$, $d = 1.97$. Though the mean accuracy for nonconflict problems was lower on this task than the base-rate task, the pattern is similar for both problems, and both the effect sizes were large.

3.3.1. Response time

There was a significant effect of accuracy on RT, $F(2, 202) = 4.76$, $p = .01$, $\eta_p^2 = 0.05$. See Table 2 for mean RT (*SD*s) on this task. Planned paired difference tests revealed that correct conflict solutions took reliably longer than incorrect conflict solutions (Bonferroni correction, $p = .01$). There was no difference between nonconflict solutions and incorrect conflict solutions. This result reflects a lack of overall conflict detection by participants at the group level.

3.3.1.1. Conflict detection index. To reiterate the crucial prediction, the three-stage model postulates a positive slope, with conflict detection increasing as accuracy on conflict problems increases. Fig. 3 shows that this relationship was essentially nonexistent, $r(100) = 0.01$, $p = .95$, one-tailed, $R^2 = 0.00004$. Additionally, the intercept was at zero, $b = 0.60$ ms, $t(100) = 0.001$, $p = .99$. This test combined with the other suggests that most participants were not detecting conflict on this task and were generally making many errors on these problems.

3.3.1.2. Cognitive decoupling index. The three-stage model showed that the relationship between the decoupling index and conflict problem accuracy should be negative: less decoupling and therefore faster processing is needed as one makes more correct responses. The observed relationship in Fig. 3 was negative, weak (underpowered), but marginal (predicted direction), $r(100) = -0.18$, $p = .07$, $R^2 = 0.03$. The regression line is the strongest trend observed across both experiments and tasks (though it is not significant). The intercept was significantly different from zero, $b = 2.59$ s, $t(101) = 2.69$, $p = .008$. Participants who answered in the belief-based direction more often tended to take longer on logic responses overall than their more accurate counterparts, as predicted by the decoupling hypothesis. However, since there were no reliable differences among the RT measures, it is difficult to state that the three-stage model is supported by the two indices.

3.3.1.3. Additional analyses. The underlying assumption for the preceding analyses was that belief-based and logic-based decisions might follow similarly to the decision made on the base-rate neglect task. Thus, the indexes from Pennycook et al. (2015) were computed for a direct comparison of the results between the two tasks. However, we acknowledge (with the help of a reviewer) that this course of action may not serve the data properly. Ultimately, we decided to keep the analysis, as it represents a behavioral exploration of the three-stage model, including statistical methodology replications. Further analyses are presented below to parse the conditional data, followed by an in-depth discussion of the findings in the context of boundary conditions and current state of the literature surrounding conditional reasoning.

As the results presented above were difficult to attribute to the three-stage model, additional analyses were performed to explore possible explanations. Rather than running the analyses as nonconflict vs. conflict problems, which ostensibly condenses the four problem types (and 16 possible combinations) into two types, a finer approach was needed. The three indices were computed on the individual argument structures (i.e., MP, MT, AC, & DA). A one-way ANOVA was performed on each index to determine if there were differences in mean RTs due to problem complexity. For the conflict detection index, there were no significant RT differences found, $F(3, 303) = 2.02$, $p = .11$, $\eta_p^2 = 0.02$. This was similar on the cognitive decoupling index, $F(3, 303) = 0.81$, $p = .49$, $\eta_p^2 = 0.01$. Regression analyses to determine RT trends across accuracy showed that none of the argument types had significant relationships between RT and accuracy. Further, a comparison of correlations between MP and MT was explored. Interestingly, MP arguments showed the predicted conflict detection effect direction (positive trend), while MT arguments showed a reverse direction (Fisher's $z = 2.14$, $p = .03$); this meant that when solving MT problems, participants spent less time on these problems when they were getting them wrong. There is some indication that conditional reasoning is more nuanced than base-rate neglect, and problem complexity has a role to play.

MP arguments mirrored the results from the base-rate neglect task, whereby believable but invalid arguments (producing conflict) are much like the base-rate problems where the stereotype is distinct and powerful, producing additional processing for what decision to make for poorer performers. More importantly, the other three argument types, which are generally viewed as more complex in conditional reasoning (e.g., De Neys, 2014), show positive trending (nonsignificant) slopes—correct answers took longer to make, especially in the case of MT problems.

3.3.2. Confidence ratings

As with the base-rate tasks, two tests of confidence ratings were performed. First, correlations (shown in Table 3) between nonconflict confidence and the conflict detection index, as well as conflict confidence and conflict detection index, were effectively zero (nonconflict: $r(100) = 0.01$, $p = .94$; conflict: $r(100) = -0.02$, $p = .81$).

The creation of the confidence rating conflict detection index and the subsequent regression test resulted in a similar null finding, $r(100) = -0.02$, $p = .86$, $R^2 < 0.001$. Participants were generally as confident in their responses on problems without conflict as they were when problems had logic and belief conflicts, even for those participants who made more logic-based responses overall.

3.4. Discussion

The goal of Experiment 2 was to extend the three-stage model's behavioral predictions to a conditional reasoning task, which also included utilizing a measure of confidence in one's answer as additional evidence for conflict detection (e.g., Frey et al., 2017; Thompson et al., 2011).

3.4.1. Base-rate neglect

On the base-rate task, the general findings of Experiment 1 were replicated: there were conflict detection and accuracy effects. However, stronger effects were observed here for some hypotheses. The conflict detection index effect was reduced, but there remained a significant relationship between conflict detection and choosing the base-rate answer on conflict problems. Additionally, there was mixed support for the cognitive decoupling index's relationship with base-rate answers on conflict problems.

[6] This process is detailed in the Supplemental materials.

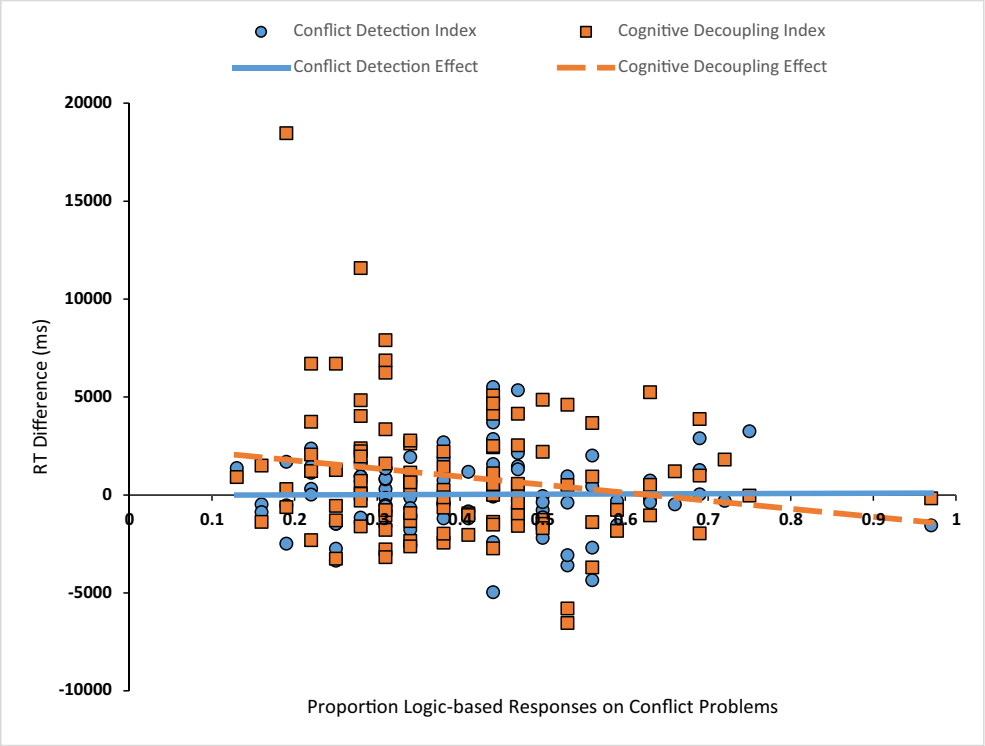

Fig. 3. Scatterplot of mean RT time differences and the proportion of correct (YES for valid and NO for invalid) responses on conflict problems the conditional reasoning task in Experiment 2. The conflict detection index refers to the RT difference between incorrect conflict responses and correct nonconflict responses. The decoupling index refers to the RT difference between correct conflict responses and overall nonconflict responses. Each unit represents one participant (i.e., one circle and square per participant). Lines show regressions of proportion of base-rate responses on RT difference scores.

As an added piece to this puzzle, confidence ratings were added as an indirect measure of monitoring and detection, based on previous work on Feeling of Rightness (FOR; Thompson, 2009; Thompson et al., 2011). However, this was not a two-response paradigm, so confidence after only one answer choice was made was assessed (e.g., Frey et al., 2017). On the base-rate neglect task, higher confidence ratings were generated more often when a person made a stereotype choice on conflict problems than when a base-rate choice was made. Furthermore, as base-rate responses increased, confidence began to waiver, indicating support for confidence as a proxy measure for conflict detection. The ratings of observed confidence match the outcomes of Thompson et al.'s (2011) Experiments 1 and 2, in which confidence differences were small but invariable. Overall, these effects reflect overconfidence in the face of conflict, a direct result of the saliency of the stereotypes on the conflict problems (Pennycook et al., 2014; Svedholm & Lindeman, 2013; Swan & Revlin, 2015), rather than the making uneasy base-rate selections.

3.4.2. Conditional reasoning

The comparison of effects from the base-rate task and conditional reasoning tasks was empirically and theoretically important to determine if the model predictions of the three-stage model apply to additional tasks (e.g., Bago & De Neys, 2017; Frey et al., 2017; Newman et al., 2017). First, the base-rate neglect task creates a conflict between the representativeness heuristic and probability principles, whereas the conditional reasoning creates a conflict between beliefs and formal logic. The purpose of comparing these two distinct tasks was not to state which task is better or more appropriately reflects human intelligence or rationality, but to determine the role of conflict detection in two qualitatively different tasks.

Creating nonconflict and conflict problems to match the structure of the base-rate task yielded a similar accuracy effect: nonconflict problems had significantly more correct answers than conflict problems. However, the replication of effects ended here. A possible explanation of these results (explored further below) could be that participants were asked if the conclusion was likely to follow the premises, rather than a directive toward the argument's validity (though, to reiterate, participants were instructed to assess whether the conclusion *logically followed* from the premises, which is a standard piece of instruction for a logical task such as conditional reasoning). This may have signaled to participants to operate pragmatically (e.g., Markovits, 2012) and interpreted the likelihood of the antecedent in relation to the consequent, regardless of the logical validity.[7] The overall large effect of nonconflict to conflict accuracy tempers this reasoning.

A logical structure analysis indicated that there were some problem complexity issues, particularly with MP and MT argument types. On the conflict detection index, MP problems replicated the base-rate neglect problems, whereas MT problems produced the opposite (nonsignificant) trend, revealing a striking difference between the two valid argument structures and the influence of believability conflict. Moreover, MT problems (along with AC and DA) showed positive trends on the inhibition index, suggesting that these more complex (and by extension, more difficult) influenced processing times when the participants made (correct) logical choices on conflict problems.

RT analyses revealed a lack of conflict detection by group or individual-level analyses; many participants were faster to get a conflict problem incorrect than make any decision on nonconflict problems. Variation among RTs on the conditional reasoning was too similar, reflecting very limited analytic engagement (as argued from the observation of the lack of seconds-level RTs). The conflict detection and cognitive decoupling indices were either not different from zero or in the wrong direction, erasing all support for the predictions of the three-stage model.

Confidence ratings were the additional measure used in this experiment to further explore conflict detection. Utilizing similar analyses from the base-rate neglect task revealed that there was no conflict detection within confidence ratings. As the RT conflict detection index was a flat relationship, so was the confidence ratings conflict index. There was invariance in responses across problem type or logical

[7] We thank an anonymous reviewer for this particular explanation, as well as suggestions from another reviewer, that instigated further analysis of the conditional reasoning data to better understand the pattern of the results.

structure.

4. General discussion

The central psychological issue of the present studies was whether or not people realize that they are biased. The preceding experiments attempted to address this central issue by investigating two major pieces within a dual process theories framework, namely (1) the role the conflict detection and monitoring mechanism, as described by Pennycook et al. (2015) in the three-stage model, and (2) and the boundary conditions of the three-stage model and if the predictions can be applied to different tasks than those from which it was originally developed.

4.1. Theoretical implications

Conflict detection and eventual resolution ought to be an essential function of any dual process framework (Bago & De Neys, 2017; De Neys, 2012, 2014; De Neys & Glumicic, 2008; Evans, 2007, 2009; Pennycook et al., 2012; Pennycook et al., 2015) because it identifies *how* and *when* a person utilizes T1 responses over T2 responses, or vice versa. Early processing models failed to account for how conflict was handled by the two processing types (e.g., Epstein, 1994; Kahneman & Frederick, 2002; Sloman, 1996), which has engendered criticism to the overall dual process framework (e.g., Keren, 2013; Kruglanski & Gigerenzer, 2011). Moreover, a specific description of the conflict mechanism is required to support the case of dual processing vs. unimodal or continuum processing (Osman, 2004). The three-stage model postulates that detection is an individual difference, and for those who utilize it, can be quite quick (Pennycook et al., 2015). Using the description of biases by De Neys and Bonnefon (2013), we address the theoretical implications of the overall results.

The results from two experiments are possibly consistent with the three-stage model, but there are a few qualifications to be made. Overall, conflict detection was faulty for those participants who tended to make many biased responses. The three-stage model is situated to account for the present data, especially on the base-rate neglect task (though the methodology was different for the task). According to the model, in Stage 1 of the thinking process, reasoners utilize T1 processing—the initial responses are generated rapidly based on intuition (Evans, 2007). If there is no conflict between various initial responses, such as when a base-rate problem cues the same answer from both the actual numbers and the stereotypic information, decisions generally proceed within milliseconds from the generation of responses. This is a good baseline for all responses, but particularly for assessing the engagement of T2. In the case of conflict in the initial responses, a resolution must be made. In this simplified dichotomy of responses, the stereotype response is at odds with the probabilistic (base-rate) response. The conflict monitoring mechanism reflects additional processing. This additional processing, whether a person decides to choose the heuristic response or the probabilistic response, reflects T2 engagement (Pennycook et al., 2015). What T2 does in the moments after the conflict is initially detected is identified by the model as cognitive decoupling or rationalization.

The three-stage model does not specifically address inhibition failure (it was classified as part of a cognitive decoupling process). The negative correlation between inhibition RTs and conflict problem responses indicated that the fewer stereotype responses made on these problems generally were generally quicker. Thus, participants with more base-rate responses or correct validity assessments generally spent less time making those responses than making stereotype or incorrect validity assessments. It makes sense that the participants who made the opposite responses had to inhibit these responses on a limited number of trials, but it is unclear if this proceeds from cognitive decoupling or rationalization within the three-stage model. This is perhaps due to the fact that the model (and the authors) make no strong claims about the time difference between decoupling and rationalization. However, the wrinkle here is that the model presented in Pennycook et al. (2015) suggests that decoupling should take longer than rationalization, especially if the assumption that stereotypes are more salient is accepted (Pennycook & Thompson, 2012), and decoupling represents a break from the first initial response (i.e., for the alternative initial response). Consequently, as rationalization represents a selection of the initial response, it should be immediately available for the decision. On the base-rate neglect task specifically, the decoupling effect did not replicate fully. However, the explanation should be tempered to show mixed support for the decoupling hypothesis; the lack of a negative correlation (as observed in Pennycook et al., 2015) does not necessarily indicate that decoupling (or other T2 processing) doesn't occur. At the very least, a limitation of the present data exists due to methodological differences and the primary predictions of the three-stage model (conflict detection) were supported.

An additional theoretical question surrounding whether people know they are making biased decisions or not is whether they have the requisite knowledge to make a fully-informed decision (i.e., avoid the biased decision by understanding all the pieces of the problem). While it is entirely plausible that some people have these *storage failures* (De Neys & Bonnefon, 2013; Stanovich, 2009), it is difficult to determine that this source of bias is behaviorally distinct from monitoring failures. Both types of failures are marked by biased responses on conflict problems and fast response times on both problem types. The data presented here do not support this distinction; storage failures were not observed. This is not surprising, as the base-rate neglect task does not necessarily lend itself to a storage failure explanation. Recent studies additionally support the notion that storage failures are an unlikely explanation and rare occurrence because T1 can cue a logical answer with high confidence (Bago & De Neys, 2017; Frey et al., 2017).

4.2. Generalization of the three-stage model

Pennycook et al. (2015) focused the model and the predictions within the base-rate neglect task. It was wholly appropriate to extend these predictions to a qualitatively different task, especially one that that wouldn't necessarily activate thinking about stereotypes or social inferences. In recent dual process theory investigations, conditional reasoning has been used (Thompson et al., 2011), as well as conjunction fallacy (De Neys et al., 2011; Frey et al., 2017), syllogisms (De Neys et al., 2010), ratio bias (Mevel et al., 2015; Thompson & Johnson, 2014), the bat and ball problem (De Neys et al., 2013; Frey et al., 2017; Johnson et al., 2016), number conservation (De Neys et al., 2014), and math reversal errors (Lubin, Simon, Houdé, & De Neys, 2015). Conditional reasoning was a reasonable choice from these tasks for replication and extension purposes. Conditional reasoning, especially inferences about definitional or causal rules, does not involve making inferences about people, merely the connection between objects or outcomes, and is therefore qualitatively distinct from base-rate neglect.

Experiment 2 contained both the base-rate neglect and conditional reasoning tasks for a direct comparison with the same participants. Perhaps the only finding that carries through both tasks is that monitoring failures are a significant source of bias. RT differences were extremely small on the conditional reasoning task, which do not match the larger effects observed on the base-rate neglect task. We are reticent to suggest that the three-stage model applies across tasks because the effects were extremely weak on the conditional reasoning task. However, some individuals do show significant conflict detection effects, which does reflect support for the three-stage model's overarching conclusion that conflict detection is an individual difference, is imperfect (some poor detectors indicated by large RT differences still answered several items correctly), and potentially inefficient across different kinds of problems.

It may be the case that the three-stage model does not readily align with the observed data because conflict detection models offer no

explanation equally. The prompt for the task used (as mentioned above), where likelihood of the conclusion following from the premises was assessed, may point to the reason why conflict detection was not detected in the task. This is not a limitation of the three-stage model, but suggests there might be different conditions in higher-order thinking tasks where detection effects might emerge. The lack of significant effects on the conditional reasoning task offers a glimpse at the overall thinking differences that result from conditional problems vs. base-rate neglect problems, as well as the way a problem is presented to participants. However, in fairness to the three-stage model, the lack of effects also do not represent a strong test of its generalizability across tasks (considering the observation of detection effects conditionals in recent studies, e.g., Newman et al., 2017), but rather a methodological nuance. These differential conditions where detection effects emerge should be explored in future research.

Are there other models that might explain the conditional reasoning data? The parallel processing model of belief bias (Handley, Newstead, & Trippas, 2011; Handley & Trippas, 2015; Trippas, Thompson, & Handley, 2017) might offer some insight, as the majority of the methodological work has utilized conditional reasoning. Recently, Trippas et al. (2017) found that MP arguments tend to be influenced more by logic than by beliefs, but MT arguments are influenced by the opposite (when participants are asked to rate the validity AND the believability of belief-bias syllogisms). These observations support the idea that beliefs and logic are concurrent instantiations, rather than serial, belief-comes-first processing. Of course, the three-stage model is designed to incorporate beliefs and normative thinking in T1 thinking; this is not the issue. Rather, the parallel processing model offers a deeper, more robust explanation of the role of conflict detection within conditional reasoning, which is not the case in base-rate neglect problems: problem complexity is a heavy influencer in the practical decisions that are made in the task. While the argument analyses in Experiment 2 did not reveal distinct, significant effects, the glimmer of differences do indicate an additional layer. The differences on the conflict detection and inhibition indices between MP and MT arguments aligns with Trippas et al.'s account rather than the logical intuitions model or the three-stage model for conditional reasoning.

It should be noted that there are some limitations to the present methodology that future research should address. These limitations may also point to a further alternative explanation for the conditional reasoning results observed in Experiment 2. Several studies on the conditional reasoning task (De Neys, Schaeken, & d'Ydewalle, 2005a, 2005b; Markovits, Fleury, Quinn, & Venet, 1998; Verschueren, Schaeken, & d'Ydewalle, 2005) and other reasoning and judgment tasks (Franssens & De Neys, 2009; Johnson et al., 2016) suggest that the complexity issue described above that it is not necessarily appropriate to assign belief-based responses to T1 or heuristic activation (initial response of the three-stage model) and logical thinking to T2 activation (alterative response). Loading working memory does not affect heuristic judgments on base-rate tasks (Franssens & De Neys, 2009; Johnson et al., 2016), but it does affect belief-based responses by reducing them on conditional reasoning problems (De Neys et al., 2005a, 2005b). The latter effect is likely due to the nature of belief-based responses: a search for counterexamples, rather than assessing validity, requires working memory and T2 engagement. Thus, belief-based responses cannot be generally assigned to T1 initial responses only. In this case, the conflict detection and cognitive decoupling indexes, based in a response time and accuracy to normative responses, are not well-suited to explore these data. As one reviewer put it: the measures are impure. This limitation might explain the lack of replication of three-stage model effects observed on the base-rate neglect. Since not all tasks are inherently equal, nor do they require the same processing, this represents a boundary condition for the three-stage model and assumptions of the model must be met before correspondence of the measures can be studied. We do not believe this limitation discounts our exploration of the generalizability of model, but it does qualify our conclusions and we do not attempt to suggest the three-stage model is to be wholly discounted because of it.

4.3. Conclusion

The purpose of the present study was to investigate the general question related to whether people know they are biased, how this bias manifests on classic reasoning and judgment tasks, and importantly, if a recent dual process theory model provides a general explanation for this conflict detection. First, conflict detection and resolution is an individual difference, as postulated by the three-stage model. Conflict monitoring is essential for detecting multiple cued responses in a given problem. Second, the model seems to have boundaries. The complexity and qualitative differences of conditional reasoning may reflect different processing from base-rate neglect (Trippas et al., 2017). For some folks, like Al Roker, it is possible that their bias is the result of a lax conflict monitoring mechanism and perhaps they do not recognize that bias.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2018.05.003.

References

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158*, 90–109. http://dx.doi.org/10.1016/j.cognition.2016.10.014.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*(3), 306–307. http://dx.doi.org/10.1207/s15327752jpa4803_13.

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science, 7*(1), 28–38. http://dx.doi.org/10.1177/1745691611429354.

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking and Reasoning, 20*(2), 169–187. http://dx.doi.org/10.1080/13546783.2013.854725.

De Neys, W., & Bonnefon, J.-F. F. (2013). The “whys” and “whens” of individual differences in thinking biases. *Trends in Cognitive Sciences, 17*(4), 172–178. http://dx.doi.org/10.1016/j.tics.2013.02.001.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*e15954. http://dx.doi.org/10.1371/journal.pone.0015954.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition, 106*(3), 1248–1299. http://dx.doi.org/10.1016/j.cognition.2007.06.002.

De Neys, W., Lubin, A., & Houdé, O. (2014). The smart non-conserver: Preschoolers detect their number conservation errors. *Child Development Research, 2014*, 1–7. http://dx.doi.org/10.1155/2014/768186.

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 208–216. http://dx.doi.org/10.3758/CABN.10.2.208.

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., Schaeken, W., & d'Ydewalle, G. (2005a). Working memory and counterexample retrieval for causal conditionals. *Thinking and Reasoning, 11*, 123–150.

De Neys, W., Schaeken, W., & d'Ydewalle, G. (2005b). Working memory and everyday conditional reasoning: Retrieval and inhibition of stored counterexamples. *Thinking and Reasoning, 11*, 349–381.

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect we are biased. *Psychological Science, 19*(5), 483–489.

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist, 49*(8), 709–724. http://dx.doi.org/10.1037//0003-066X.49.8.709.

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454–459. http://dx.doi.org/10.1016/j.tics.2003.08.012.

Evans, J. S. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking and Reasoning, 13*(4), 321–339. http://dx.doi.org/10.1080/13546780601008825.

Evans, J. S. B. T. (2009). How many dual-process theories do we need? One, two, or many? In J. S. B. T. Evans, & K. Frankish (Eds.). *In two minds: Dual process and beyond* (pp. 33–54). Oxford: Oxford University Press.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241. http://dx.doi.org/10.1177/1745691612460685.

Frankish, K., & Evans, J. S. B. T. (2009). The duality of mind: An historical perspective. In J. S. B. T. Evans, & K. Frankish (Eds.). *In two minds: Dual process and beyond* (pp. 1–29). Oxford: Oxford University Press.

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking and Reasoning, 15*, 105–128.
Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19*(4), 25–42. http://dx.doi.org/10.1257/089533005775196732.
Frey, D., Johnson, E. D., & De Neys, W. (2017). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology,* 1–52. http://dx.doi.org/10.1080/17470218.2017.1313283.
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*(1), 28–43.
Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. In B. H. Ross (Vol. Ed.), *The psychology of learning and motivation. Vol. 62. The psychology of learning and motivation* (pp. 33–58). Waltham, MA: Academic Press. http://dx.doi.org/10.1016/bs.plm.2014.09.002.
Johnson, E. D., Tubau, E., & De Neys, W. (2016). The doubting system 1: Evidence for automatic substitution sensitivity. *Acta Psychologica, 164*, 56–64.
Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.). *Heuristics and biases: The psychology of intuitive judgments* (pp. 49–81). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511808098.
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*(3), 257–263. http://dx.doi.org/10.1177/1745691613483474.
Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97–109. http://dx.doi.org/10.1037/a0020762.
Lubin, A., Houdé, O., & De Neys, W. (2015). Evidence for children's error sensitivity during arithmetic word problem solving. *Learning and Instruction, 40*, 1–8.
Lubin, A., Simon, G., Houdé, O., & De Neys, W. (2015). Inhibition, conflict detection, and number conservation. *ZDM Mathematics Education, 47*, 793–800.
Markovits, H. (2012). Pragmatic reasoning schemas. In N. M. Seel (Ed.). *Encyclopedia of the sciences of learning* (pp. 2660–2662). US: Springer.
Markovits, H., Fleury, M. L., Quinn, S., & Venet, M. (1998). Conditional reasoning and the structure of semantic memory. *Child Development, 69*, 742–755.
Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology, 27*(2), 227–237.
Newman, I. R., Gibb, M., & Thompson, V. A. (2017). Rule-based reasoning is fast and belief-based reasoning can be slow: Challenging current explanations of belief-bias and base-rate neglect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(7), 1154–1170. http://dx.doi.org/10.1037/xlm0000372.
Oliver, J., & Pennolino, P. (2016). Scientific studies [television series episode]. In T. Carvell, J. Oliver, J. Taylor, & J. Thoday (Eds.). *Last Week Tonight with John Oliver*. New York: Home Box Office, Inc.
Osman, M. (2004). An evaluation of dual-process theories of reasoning. *Psychonomic Bulletin & Review, 11*(6), 988–1010.
Pennycook, G. (2017). A perspective on the theoretical foundation of dual process models. In W. De Neys (Ed.). *Current issues in thinking and reasoning: Dual process models 2.0*. New York: Psychology Press.
Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition, 124*(1), 101–106. http://dx.doi.org/10.1016/j.cognition.2012.04.004.
Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80*(October), 34–72. http://dx.doi.org/10.1016/j.cogpsych.2015.05.001.
Pennycook, G., & Thompson, V. A. (2012). Reasoning with base rates is routine, relatively effortless, and context dependent. *Psychonomic Bulletin & Review, 19*(3), 528–534. http://dx.doi.org/10.3758/s13423-012-0249-3.
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*(2), 544–554. http://dx.doi.org/10.1037/a0034887.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22. http://dx.doi.org/10.1037//0033-2909.119.1.3.
Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. S. B. T. Evans, & K. Frankish (Eds.). *In two minds: Dual process and beyond* (pp. 55–88). Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199230167.003.0003.
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*(2), 342–357.
Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23*(5), 645–665 (discussion 665–726).
Svedholm, A. M., & Lindeman, M. (2013). The separate roles of the reflective mind and involuntary inhibitory control in gatekeeping paranormal beliefs and the underlying intuitive confusions. *British Journal of Psychology, 104*, 303–319. http://dx.doi.org/10.1111/j.2044-8295.2012.02118.x.
Swan, A. B., & Revlin, R. (2015). Inhibition failure is mediated by a disposition toward flexible thinking. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.). *Proceedings of 37th Annual Meeting of the Cognitive Science Society* (pp. 2314–2319). Austin, TX: Cognitive Science Society.
Thompson, V. A. (1994). Interpretational factors in conditional reasoning. *Memory & Cognition, 22*(6), 742–758. http://dx.doi.org/10.3758/BF03209259.
Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. In J. S. B. T. Evans, & K. Frankish (Eds.). *In Two Minds: Dual Process and Beyond* (pp. 171–195). Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780199230167.003.0008.
Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning, 20*, 215–244. http://dx.doi.org/10.1080/13546783.2013.869763.
Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107–140. http://dx.doi.org/10.1016/j.cogpsych.2011.06.001.
Thompson, V. A., Prowse Turner, J. A., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*(2), 237–251. http://dx.doi.org/10.1016/j.cognition.2012.09.012.
Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory and Cognition, 45*, 539–552.
Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131.
Verschueren, N., Schaeken, W., & d'Ydewalle, G. (2005). A dual process specification of causal conditional reasoning. *Thinking and Reasoning, 11*, 239–278.