The Reflective Mind: Examining Individual Differences in Susceptibility to Base Rate Neglect with fMRI

Oshin Vartanian[1,2], Erin L. Beatty[1,3], Ingrid Smith[1], Kristen Blackler[1], Quan Lam[1], Sarah Forbes[1,4], and Wim De Neys[5,6,7]

Abstract

■ Performance on heuristics and bias tasks has been shown to be susceptible to bias. In turn, susceptibility to bias varies as a function of individual differences in cognitive abilities (e.g., intelligence) and thinking styles (e.g., propensity for reflection). Using a classic task (i.e., lawyer–engineer problem), we conducted two experiments to examine the differential contributions of cognitive abilities versus thinking styles to performance. The results of Experiment 1 demonstrated that the Cognitive Reflection Test (CRT)—a well-established measure of reflective thinking—predicted performance on conflict problems (where base rates and intuition point in opposite directions), whereas STM predicted performance on nonconflict problems. Experiment 2 conducted in the fMRI scanner replicated this behavioral dissociation and enabled us to probe their neural correlates. As predicted, conflict problems were associated with greater activation in the ACC—a key region for conflict detection—even in cases when participants responded stereotypically. In participants with higher CRT scores, conflict problems were associated with greater activation in the posterior cingulate cortex (PCC), and activation in PCC covaried in relation to CRT scores during conflict problems. Also, CRT scores predicted activation in PCC in conflict problems (over and above nonconflict problems). Our results suggest that individual differences in reflective thinking as measured by CRT are related to brain activation in PCC—a region involved in regulating attention between external and internal foci. We discuss the implications of our findings in terms of PCC's possible involvement in switching from intuitive to analytic mode of thought. ■

INTRODUCTION

A major theme emerging from the heuristics and bias literature has been that human reasoning can at times deviate from normative principles of statistical prediction. One task that has been used repeatedly to demonstrate such an effect is Kahneman and Tversky's (1973) classic lawyer–engineer problem. In the original version of the task, participants were presented with the following description:

> A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

One of the five descriptions provided read as follows:

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include some carpentry, sailing, and mathematical puzzles. The probability that Jack is one of the 30 engineers in the sample of 100 is ___%.

Despite the fact that, based on the prior odds, a randomly selected person would most likely be a lawyer, participants' subjective probabilities revealed that they were more likely to consider Jack to be an engineer. Kahneman and Tversky (1973) explained this discrepancy in terms of the representativeness heuristic, according to which base rates are largely ignored if individuating information is made available, which in turn drives people to respond stereotypically based on the "essential features of the evidence" (p. 238). Although this bias has been replicated many times, the precise reasons underlying it continue to be a matter of debate. One influential model used to explain it is dual-process theory, according to which base rate neglect occurs because the task creates a conflict between an intuitively cued response consistent with the individuating information and an analytically cued response consistent

[1]Defence Research and Development Canada, Toronto Research Centre, [2]University of Toronto, [3]University of Southern Denmark, [4]University of Waterloo, [5]Paris Descartes University, [6]Caen Basse-Normandie University, [7]CNRS, UMR 8240, LaPsyDÉ, France

with the base rate information. More specifically, based on the default interventionist variant of dual-process theory (Evans & Stanovich, 2013; Kahneman, 2011), the lawyer–engineer problem results in a fast and automatic intuitive response, which unless curtailed will be normatively incorrect (i.e., stereotypical). Its correction depends on the downstream involvement of the analytic system (see also Kahneman, 2000).

However, there are two lines of evidence that cast doubt on the default interventionist account of base rate neglect. First, there is behavioral evidence to suggest that, despite responding stereotypically, biased reasoners are nevertheless sensitive to the presence of conflict in the form of increased response latency (e.g., Stupple, Ball, & Ellis, 2013; Bonner & Newell, 2010; Villejoubert, 2009; De Neys & Glumicic, 2008) and decreased post-decisional confidence (e.g., Gangemi, Bourgeois-Gironde, & Mancini, 2015; De Neys, Rossi, & Houdé, 2013; De Neys, Cromheeke, & Osman, 2011). Second, based on a series of experiments using the two-response paradigm,[1] it has been demonstrated that participants can generate initial responses in accordance with base rates quickly and with high degree of confidence (Bago & De Neys, 2017; Newman, Gibb, & Thompson, 2017). In other words, people can respond logically to conflict problems in an intuitive manner. This finding is consistent with the idea that some elementary logical processing might be occurring relatively early in the reasoning process (e.g., De Neys, 2012, 2014; see also Pennycook, Fugelsang, & Koehler, 2015). Overall, these results suggest that there are multiple paths for arriving at the correct solution to a base rate problem—involving both deliberate and intuitive processes.

Of particular relevance to this study, De Neys, Vartanian, and Goel (2008) administered a variety of base rate problems to participants in the fMRI scanner. On conflict problems, base rates and intuition pointed in opposite directions, whereas on nonconflict problems base rates and intuition pointed in the same direction. As expected, conflict problems were associated with greater activation in the ACC—a key region for conflict detection (see Prado & Noveck, 2007; van Veen & Carter, 2006; Botvinick, Cohen, & Carter, 2004; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuis, 2004). Interestingly, related electrophysiological work has demonstrated that participants are attuned not only to the presence of conflict but also to its degree (Prado, Kaliuzhna, Cheylus, & Noveck, 2008). In addition, when conflict problems were solved correctly, this resulted in additional activation in the right inferior frontal gyrus (IFG)—a region of the brain involved in response inhibition (see Aron, Robbins, & Poldrack, 2004, 2014; Aron, Fletcher, Bullmore, Sahakian, & Robbins, 2003; see also Prado & Noveck, 2007). This finding is consistent with the idea that responding in accordance with base rates on conflict problems necessitates not only the detection of conflict but also the deployment of executive resources for suppressing a prepotent heuristic response

in favor of a normative response. However and perhaps most interestingly, ACC was also activated when participants responded stereotypically to conflict problems, suggesting that there was an awareness of bias despite the inability to overcome it (see also Simon, Lubin, Houdé, & De Neys, 2015).

Individual Differences and Base Rate Neglect

One approach to elucidating the mechanisms that underlie performance impairments in heuristics and bias tasks has been to focus on theoretically relevant individual differences that predict performance. In an early contribution to this line of research, Stanovich and West (1998) examined the relationship between cognitive ability and performance on a host of tasks in this domain. Their work was motivated by the premise that "more reflective and engaged reasoners will be more likely to affirm the axioms that define normative reasoning" (p. 293; see also Slovic & Tversky, 1974). They operationalized cognitive ability based on Scholastic Aptitude Test (SAT) scores because they are known to gauge intellectual engagement, reflective thought, and thorough information processing. As predicted, participants with higher SAT scores were less likely to fall prey to deviations from normative axioms than participants lower in cognitive ability. Importantly, the effect of cognitive ability was greater if the problem engaged both analytic and heuristic modes of information processing that cued opposite responses. In such cases, the positive contribution of higher SAT scores was attributed to deeper engagement with the problems, in turn leading to better performance (see also Stanovich & West, 2000, 2008).

Unlike SAT scores that represent a blend of cognitive ability and thinking style, more recent work has systematically parsed out the relative contributions of individual differences in cognitive abilities (e.g., intelligence) and thinking styles (e.g., propensity for reflection) on performance in heuristics and biases tasks. Toplak, West, and Stanovich (2011) examined the contributions of cognitive ability and thinking style on performance on a wide host of tasks involving probabilistic reasoning, hypothetical thought, theory justification, scientific reasoning, and the tendency to think statistically. Cognitive ability was measured using the Vocabulary and Matrix Reasoning subtests from the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999), and the disposition to think analytically was measured using the Cognitive Reflection Test (CRT; Frederick, 2005)—an instrument designed to measure the propensity to suppress a fast intuitive response in favor of a reflective, deliberative response (e.g., A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost? Incorrect answer = 10 cents; correct answer = 5 cents). CRT problems are believed to be ideal for probing the interplay between heuristic and analytic thinking

precisely because the incorrect but intuitive response typically appears immediately, which in turn must be suppressed in favor of the correct and analytically derived response. Toplak et al. (2011) demonstrated that CRT scores were a unique predictor of performance on heuristics and biases tasks after controlling for cognitive ability. Echoing the findings of Toplak et al. (2011), Pennycook, Cheyne, Barr, Koehler, and Fugelsang (2014) have demonstrated that CRT scores and scores on a brief measure of verbal intelligence (WordSum) were both independent predictors of conflict base rate problems. The results of Toplak et al. (2011) and Pennycook, Cheyne, et al. (2014) suggest that both cognitive ability and thinking styles contribute positively to performance on base rate problems, although CRT scores appear to be a stronger predictor of conflict detection.

The Present Study

Although previous research has already examined the neural bases of choices made under conflict in the context of heuristics and biases and reasoning tasks (e.g., De Neys et al., 2008; Prado et al., 2008; Prado & Noveck, 2007), there is a dearth of research that has explicitly focused on parsing out the relative contributions of cognitive abilities and thinking styles in relation to performance on those tasks involving neuroimaging data. To address that gap, we conducted two experiments to assess the relative contributions of cognitive ability (intelligence and STM) and thinking style (CRT) to performance on conflict and nonconflict base rate problems. In our behavioral study (Experiment 1), all measures were administered in our laboratory using paper-and-pencil format, and performance on the base rate task was not timed. In contrast, in the second study, the base rate task was administered inside the fMRI scanner and was timed. Our design enabled us to (a) expand our measures of cognitive ability beyond intelligence to also include measures of STM and (b) assess the reliability of our behavioral findings across two experiments. Critically, however, our design enabled us to probe the neural correlates of individual differences in CRT scores while participants were engaged in conflict problems. We hypothesized that (a) CRT scores would predict performance uniquely on conflict but not on nonconflict problems and (b) CRT scores would be related to brain activation to a greater extent during conflict than nonconflict problems. Given that CRT is believed to reflect one's ability to inhibit an intuitive response and switch to an analytic mode of thought, we expected that CRT scores would covary with brain activation in brain regions associated with executive functions (Miyake et al., 2000)—in particular, right IFG given its role in inhibition (see Aron et al., 2003, 2004, 2014) and the pFC and ACC given their contributions to set shifting (see Bissonette, Powell, & Roesch, 2013; Robbins, 1996).

EXPERIMENT 1

Method

Participants

The participants were 30 volunteers (25 men, 5 women) recruited from the Department of National Defence. The participants ranged in age from 20 to 46 years ($M = 29.4 \pm 6.5$ years). Their levels of education were as follows: high school diploma ($n = 9$), college diploma ($n = 7$), university degree ($n = 10$), graduate degree ($n = 3$), none of the above ($n = 1$). The protocol for the study was approved by the Defence Research and Development Canada (DRDC) Human Research Ethics Committee.

Materials and Procedures

All measures were administered in our laboratory at DRDC (Toronto Research Centre). We administered two STM tasks, modeled after Harrison et al.'s (2013) simple working memory span tasks. For word (verbal) span, four-letter monosyllabic words were presented one at a time on a monitor. After each block of words, participants were prompted by the software to recall the words they saw in the order in which they were presented. Blocks ranged from three to nine words. For matrix span, participants were presented with a 4×4 matrix where one square (out of 16) appeared in red and the rest in white. At the end of each block of matrices, participants were instructed to recall the locations of the red squares in the order in which they were presented. Blocks ranged from three to nine matrices. The computer task provided a detailed description of each task before the start, and the experimenter reviewed the instructions and provided an example in each case to the participants. Note that both the word and matrix span are so-called simple working memory span tasks that primarily tax STM storage capacity (e.g., Harrison et al., 2013; Cowan, 2008; Unsworth & Engle, 2007).

Our measures of crystallized and fluid intelligence consisted of the Vocabulary (10 min) and Block Patterns (10 min) subsets of the Shipley-2, which were in turn standardized and converted into a single full-scale intelligence score (Shipley, Gruber, Martin, & Klein, 2009). We administered the seven-item version of the CRT, which built on Frederick's (2005) original three-item version by adding four more items (Toplak, West, & Stanovich, 2014).

Our 48 base rate problems (24 conflict, 24 nonconflict) were selected from Pennycook, Cheyne, et al.'s (2014) item pool. The participants were informed that, 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 populations groups (e.g., carpenters and policemen). In each study, one participant was drawn at random from the sample. They were informed that they would get to see a personality trait for this randomly chosen participant and that they would also receive information about the composition of the population groups

Table 1. The Descriptive Statistics for Experiment 1 and Experiment 2

	Conflict	Nonconflict	IQ	CRT	WS	MS
Exp. 1	57.73% (41.09)	97.93% (5.97)	25.07 (3.30)	53.56% (29.81)	5.58 (1.30)	5.61 (.93)
Exp. 2	51.36% (12.51)	95.44% (4.63)	25.83 (3.90)	53.79% (34.17)	9.49 (1.90)	8.75 (2.09)

Standard deviations appear in parentheses. Exp. = Experiment; IQ = intelligence as measured by Shipley-2; MS = matrix span; WS = word span.

tested in the study in question. For each problem, they would then be asked to indicate to which group the participant most likely belonged. They were then given two practice problems for familiarization purposes. The following depicts a representative item from the 48-item set:

> This study contains:
> Lawyers and clowns
> Person 'L' is argumentative
> There are 3 lawyers/997 clowns
> Person 'L' is more likely to be:
> 1) Clown
> 2) Lawyer

Note that on all problems, the base rate contrast between the two categories was similarly extreme. Two randomized orders of the 48-item set were prepared. Each problem appeared on a separate sheet of paper in a booklet. The two sets were administered randomly to the participants. There was no time limit for completing the task.

Results

Descriptive statistics and zero-order correlations for all variables are reported in Tables 1 and 2. As expected, accuracy was significantly higher for nonconflict than conflict problems, $t(29) = 5.45$, $p < .001$, $d = 1.27$.[2] In

Table 2. The Zero-order Correlation Matrix for All Variables in Experiment 1 and Experiment 2

	Conflict	Nonconflict	IQ	CRT	WS	MS
Conflict	–	.19	.30	.46*	.32	.22
Nonconflict	.34*	–	.30	.35	.09	.39*
IQ	.29	.35*	–	.44*	.40*	.49**
CRT	.41**	.42**	.52***	–	.43*	.46*
WS	.32*	.44**	.50***	.43**	–	.55**
MS	.22	.36*	.46**	.57***	.53***	–

Experiment 1 = above the diagonal; Experiment 2 = below the diagonal; IQ = intelligence as measured by Shipley-2; MS = matrix span; WS = word span.

*$p < .05$.

**$p < .01$.

***$p < .001$.

addition, participants did not always exhibit internal consistency in their response pattern to conflict problems: Whereas 15 participants responded consistently in accordance with base rates or stereotypically, the remaining 15 participants registered both types of responses across conflict problems. Next, using step-wise regression, we examined the effects of STM (verbal, matrix), intelligence, and CRT scores—separately for conflict and nonconflict problems. Matrix span was the only significant predictor of performance on nonconflict problems, $\beta = .026$, $p = .041$, accounting for 15% of the variance (R^2) in performance. In contrast, CRT was the only significant predictor of performance on conflict problems, $\beta = .607$, $p = .018$, accounting for 21.6% of the variance (R^2) in performance. Because of statistically significant correlations among our predictors, we have also reported the results of regressions involving accuracy on conflict and nonconflict problems separately for each predictor (Table 3). As was the case before, CRT was the sole predictor of performance on conflict problems, whereas matrix span was the sole predictor of performance on nonconflict problems.

Table 3. Regressions of Accuracy Involving Conflict and Nonconflict Problems on Individual Predictors in Experiment 1 and Experiment 2

Variable	Conflict β	Nonconflict β
Experiment 1		
CRT	.62 (.24)*	.07 (.04)
IQ	.03 (.02)	.01 (.00)
WS	.10 (.06)	.00 (.01)
MS	.01 (.09)	.03 (.01)*
Experiment 2		
CRT	.00 (.00)**	.00 (.00)**
IQ	.01 (.01)	.00 (.00)*
WS	.02 (.01)*	.01 (.00)**
MS	.01 (.01)	.01 (.00)*

Standard errors of measurement appear in parentheses. β = Unstandardized weights; IQ = intelligence as measured by Shipley-2; WS = word span; MS = matrix span.

*$p < .05$.

**$p < .01$.

EXPERIMENT 2

Method

Participants

The participants were 44 neurologically healthy right-handed volunteers (31 men, 13 women) with normal or corrected-to-normal vision recruited from the Department of National Defence. No participant reported color blindness. The participants ranged in age from 20 to 56 years ($M = 35.5 \pm 11.3$ years). Their levels of education were as follows: high school diploma ($n = 12$), college diploma ($n = 6$), university degree ($n = 19$), graduate degree ($n = 7$), none of the above ($n = 2$). The protocol for the study was approved by the DRDC Human Research Ethics Committee and by the Research Ethics Board of Sunnybrook Health Sciences Centre. Handedness was assessed using a standard self-report questionnaire (Oldfield, 1971).

Materials and Procedure

With the exception of the base rate task, all measures were administered in our laboratory at DRDC (Toronto Research Centre). Shipley-2 and STM tasks were identical to Experiment 1. In contrast, we divided the CRT randomly into three-item and four-item versions, and each participant completed one of the two versions. The reason for the discrepancy in the number of CRT items between Experiment 1 and Experiment 2 was that the latter was conducted as part of a larger study for which many other measures unrelated to this specific experiment were also administered, forcing us to minimize the duration of each individual paper-and-pencil measure. However, critically, note that average accuracy rate and variance on CRT scores in Experiment 2 is essentially identical to Experiment 1:

53.56% ($SD = 29.81$, range $= 0–100$). The base rate task included the exact same 48 items as the set administered in Experiment 1 (Pennycook, Cheyne, et al., 2014) but, in this case, was administered inside the fMRI scanner (Figure 1). Before entering the scanner, participants were given two practice problems for familiarization purposes.

Image Acquisition and Processing

A 3-T MR scanner with an eight-channel head coil (Discovery MR750, 22.0 software, GE Healthcare) was used to acquire T1 anatomical volume images ($0.86 \times 0.86 \times 1.0$ mm voxels). For functional imaging, T2*-weighted gradient-echo spiral-in/out acquisitions were used to produce 26 contiguous 5-mm thick axial slices (repetition time $= 2000$ msec, echo time $= 30$ msec, flip angle $= 70°$, field of view $= 200$ mm, 64×64 matrix, voxel dimensions $= 3.1 \times 3.1 \times 5.0$ mm), positioned to cover the whole brain. The first five volumes were discarded to allow for T1 equilibration effects. The number of volumes acquired was 303.

Statistical Analysis

Data were analyzed using Statistical Parametric Mapping (SPM8; www.fil.ion.ucl.ac.uk/spm/). Head movement was less than 2 mm in all cases. All functional volumes were spatially realigned to the first volume. Given that the volumes were acquired using a descending sequence with short repetition time, slice timing to correct for variation in acquisition time followed realignment (Huettel, Song, & McCarthy, 2004). A mean image created from realigned volumes was spatially normalized to the MNI EPI brain template using nonlinear basis functions. The derived spatial transformation was applied to the realigned T2*

Figure 1. Procedure for the base rate task. Adapted with permission from Pennycook, Cheyne, et al. (2014).
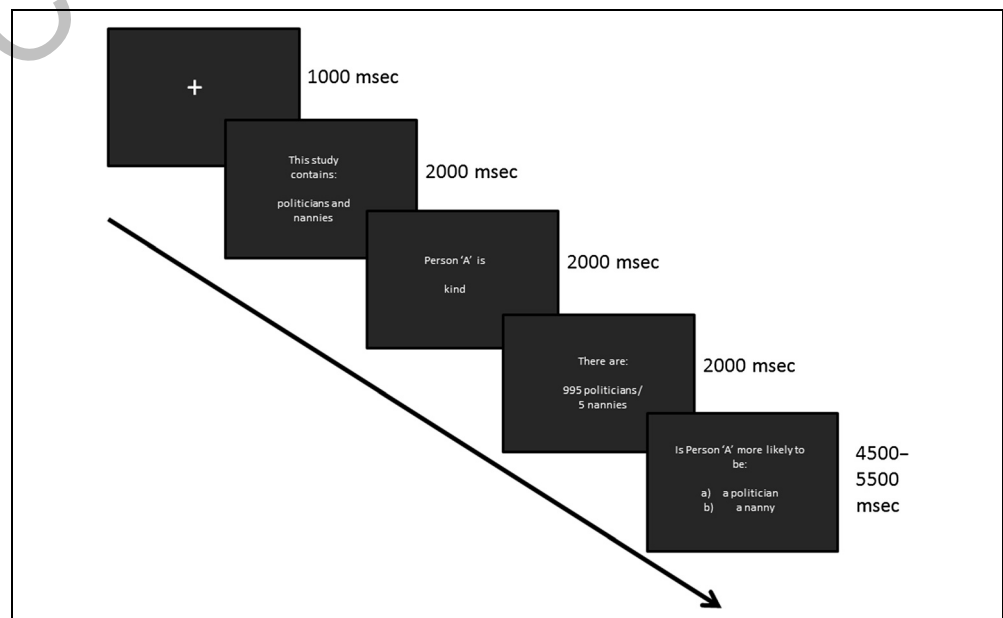
volumes and spatially smoothed with an 8-mm FWHM isotropic Gaussian kernel. Time series across each voxel were high-pass filtered with a cutoff of 128 sec, using cosine functions to remove section-specific low-frequency drifts in the BOLD signal. Condition effects at each voxel were estimated according to the general linear model, and regionally specific effects were compared using linear contrasts. The BOLD signal was modeled as a box-car, convolved with a canonical hemodynamic response function. We applied a cluster-level correction within SPM8 for determining statistical significance. Specifically, reported activations survived a voxel-level threshold of $p < .001$ (uncorrected for multiple comparisons) and a cluster-level threshold of $p < .05$ corrected for multiple comparisons (FWE).

Using an event-related design, in the first level, we specified regressors corresponding to the following points in the problem structure (see Figure 1): (1) fixation point, (2) the groups in question, (3) stereotype information, (4) base rates, (5) prompt, and (6) motor response. In addition, the RT associated with each motor response was included in the model as a parameter and modeled out of the analyses by assigning a value of 0 to its regressor in subsequent analyses. All reported neural analyses are based on the prompt time point (i.e., last slide in Figure 1). Importantly, prompt problems were in turn separated into four separate regressors based on performance as follows: conflict (correct), conflict (incorrect), nonconflict (correct), nonconflict (incorrect).

Results

Behavioral

Descriptive statistics and zero-order correlations for all variables are reported in Tables 1 and 2. As expected, accuracy was significantly higher for nonconflict than conflict problems, $t(43) = 24.84$, $p < .001$, $d = 4.40$. Here, too, the participants were not internally consistent in terms of their response pattern to conflict problems. Specifically, no participant responded consistently in accordance with base rates or stereotypically across conflict problems. In addition, in terms of RT, we replicated Pennycook, Cheyne, et al.'s (2014) result by demonstrating that RT was significantly longer for conflict problems regardless of stereotypical or base rate responses compared with correct responses on nonconflict problems, $F(2, 86) = 16.21$, $p < .001$, $\eta_p^2 = .27$ (Figure 2).[3] Next, using step-wise regression, we examined the effects of STM (verbal, matrix), intelligence, and CRT scores—separately for conflict and nonconflict problems. Word span was the only significant predictor of performance on nonconflict problems, $\beta = .011$, $p = .003$, accounting for 19.2% of the variance (R^2) in performance. In contrast, CRT was the only significant predictor of performance on conflict problems, $\beta = .001$, $p = .006$, accounting for 16.5% of the variance (R^2) in performance. As was
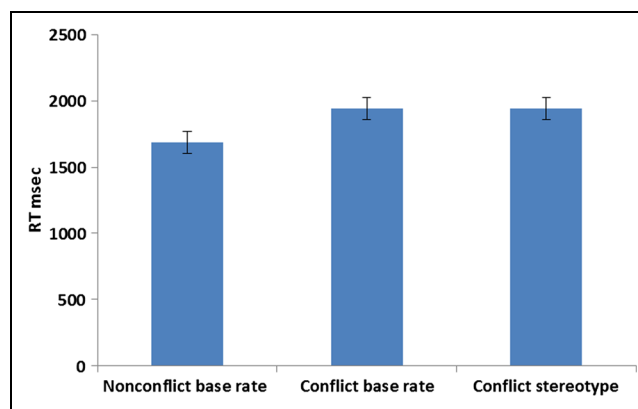

Figure 2. RT as a function of conflict in Experiment 2. Nonconflict base rate = base rate response on nonconflict problems; Conflict base rate = base rate response on conflict problems; Conflict stereotype = stereotypical response on conflict problems. Error bars represent *SEM*.

the case in Experiment 1, because here too we observed statistically significant correlations among our predictors, we have reported the results of regressions involving accuracy on conflict and nonconflict problems separately for each predictor (Table 3). As was the case before, CRT emerged as the strongest predictor of performance on conflict problems, but word span was also a significant predictor. In contrast, individually, all four independent variables emerged as significant predictors to performance on nonconflict problems.

Neural

We began our analysis by first attempting to replicate the basic findings reported in De Neys et al. (2008), despite some differences between the designs of the two studies. First, De Neys et al. (2008) reported that there was greater activation in ACC ($x = -2$, $y = 24$, $z = 44$) when they compared conflict to nonconflict problems, which was attributed to the detection of conflict in the former condition. Our comparison of those same conditions here also revealed relatively greater activation in an identical location within a cluster that included ACC bilaterally ($x = -4$, $y = 24$, $z = 44$; $Z = 3.84$, BA 32; $x = 2$, $y = 28$, $z = 38$; $Z = 3.79$, BA 32), as well as the left superior frontal gyrus ($x = -14$, $y = 46$, $z = 44$; $Z = 4.18$, BA 8; Figure 3), and in a separate cluster overlapping with the left middle temporal gyrus ($x = -48$, $y = -70$, $z = 22$; $Z = 4.33$, BA 39). To observe whether the effect observed in ACC was due to an increase in brain activation in conflict problems (rather than a decrease in brain activation in nonconflict problems), we used the MarsBaR toolbox (Brett, Anton, Valabregue, & Poline, 2002) implemented in SPM8 to calculate parameter estimates centered around ACC as our ROI (center of mass: $x = -2$, $y = 24$, $z = 44$; radius = 10 mm)—separately for conflict and nonconflict problems. The results illustrate that the
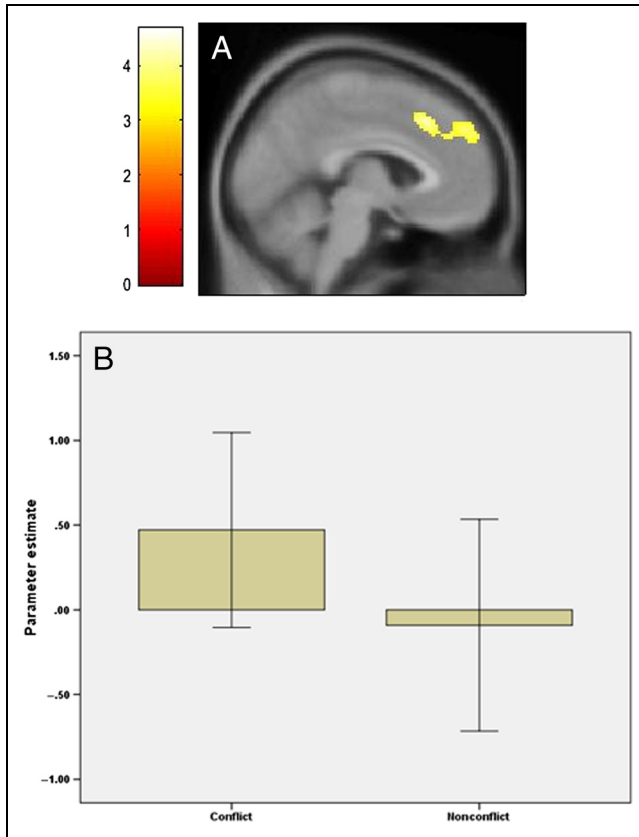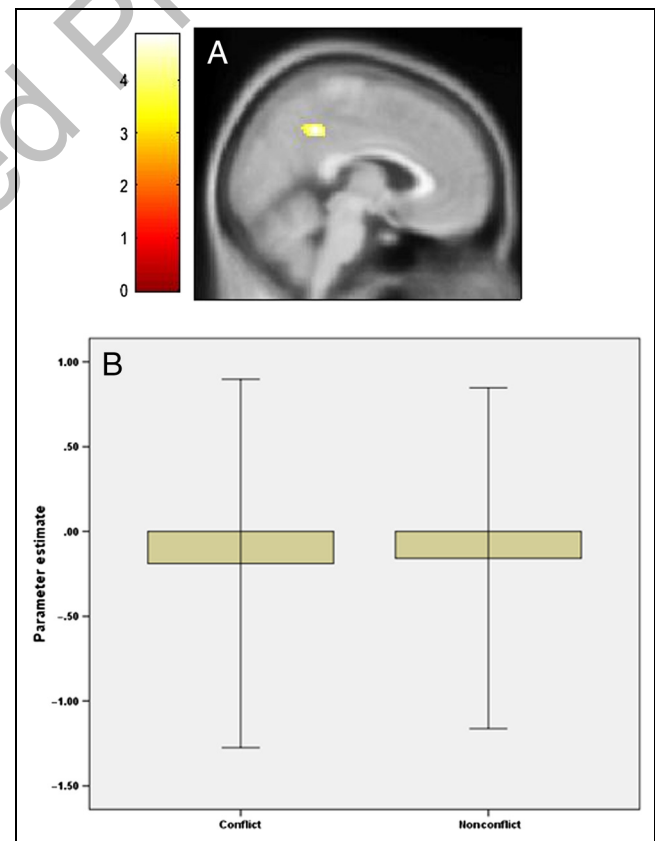
Figure 3. Conflict and activation in the ACC. (A) There is relatively greater activation in ACC in conflict versus nonconflict problems. SPM rendered into standard stereotactic space and superimposed onto sagittal MRI in standard space. Bar represents the corresponding *t* score. (B) The parameter estimates illustrate that the observed effect was due to an increase in brain activation in ACC in conflict problems. Error bars represent 95% confidence intervals.

effect observed in ACC was due to an increase in brain activation in conflict problems (Figure 3).

Next, and critically, De Neys et al. (2008) were also able to show that, even when participants did not generate the base rate response on incongruent problems, they nevertheless activated ACC. This was interpreted to mean that they were aware of their biased responding. In the present experiment, we could not reproduce the identical contrasts because our design did not include neutral control items (in which the individuating descriptions were completely neutral) or heuristic control items (in which the base rates were neutral). Instead, we contrasted conflict problems with an incorrect response versus nonconflict problems with a correct response. In both cases, the response would be the same (i.e., consistent with base rate), but only in the former type of problems would a conflict be elicited. Indeed, the results demonstrated relatively greater activation in the former condition in a cluster that included ACC bilaterally ($x = -4, y = 24, z = 44; Z = 3.79$, BA 32; $x = 4, y = 30, z = 36; Z = 4.27$, BA 32), as well as the left superior frontal gyrus ($x = -6, y = 48, z = 36; Z = 4.05$, BA 8), and two

other clusters overlapping with the left insula ($x = -38, y = 26, z = 4; Z = 4.77$, BA 13) and the precuneus ($x = 2, y = -56, z = 36; Z = 4.30$, BA 7). Finally, De Neys et al. (2008) had also reported that right IFG was activated more on conflict problems wherein participants had generated base rate rather than stereotypical responses. This was interpreted to mean that the generation of the base rate response involved inhibition of an intuitive response. That comparison did not result in any significant activation in this study.

Next, we took two approaches for testing our focal hypotheses (i.e., that CRT scores would predict performance uniquely on conflict but not on nonconflict problems and that CRT scores would be related to brain activation to a greater extent during conflict than nonconflict problems). We reasoned that convergent findings across two approaches would reinforce the reliability of our results. First, based on a median split, we divided our sample into participants who scored high ($M = 84\%, SD = 15$) versus low ($M = 24\%, SD = 19$) on the CRT, $t(42) = 11.69, p < .001, d = 3.54$. Next, an

Figure 4. Cognitive reflection and activation in the PCC. (A) There was relatively greater activation in PCC in participants with higher compared with lower CRT scores in relation to conflict versus nonconflict problems. SPM rendered into standard stereotactic space and superimposed onto sagittal MRI in standard space. Bar represents the corresponding *t* score. (B) The parameter estimates illustrate that, across all participants, activation in PCC did not vary as a function of conflict. Error bars represent 95% confidence intervals.
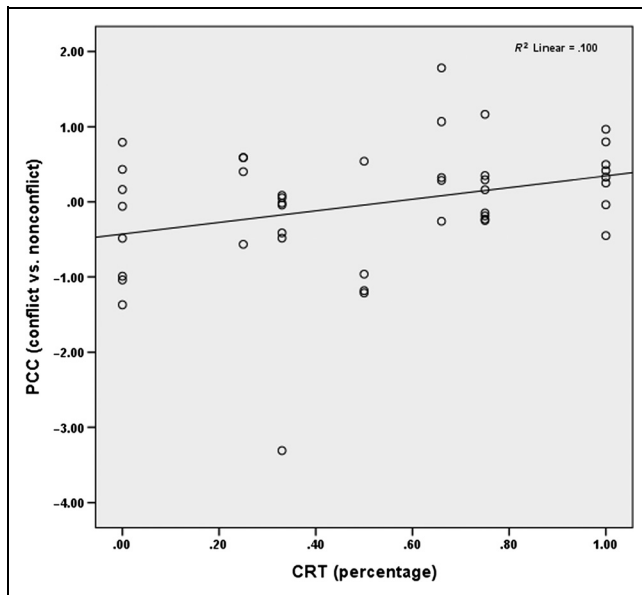
Figure 5. CRT scores predict brain activation in PCC involving conflict.

independent-samples *t* test involving the fMRI data demonstrated that, compared with participants with low CRT scores, in participants with higher CRT scores conflict (compared with nonconflict) problems were associated with greater activation in the posterior cingulate cortex (PCC; $x = 4$, $y = -44$, $z = 40$; $Z = 4.31$, BA 31; Figure 4). Second, we explored brain regions where activation would covary as a function of individual differences in CRT scores in relation to conflict (compared with nonconflict) problems. Converging on a similar result as before, the PCC was the only brain region where activation as a function of CRT scores covaried in relation to conflict (compared with nonconflict) problems ($x = 4$, $y = -48$, $z = 40$; $Z = 3.71$, BA 31).[4] Next, we examined whether individual differences in CRT scores would predict brain activation in the PCC in conflict (vs. nonconflict) contrast. Specifically, we used the MarsBaR toolbox (Brett et al., 2002) implemented in SPM8 to calculate parameter estimates centered around the PCC as our ROI (center of mass: $x = 4$, $y = -44$, $z = 40$; radius = 10 mm) for the conflict versus nonconflict contrast. The results revealed that CRT scores were a significant predictor of activation in PCC, $\beta = .78$, $t = 2.17$, $p < .05$ (Figure 5). Indeed, CRT scores accounted for 10% of the variance in PCC activation in relation to the conflict problems versus nonconflict problems.

GENERAL DISCUSSION

This study was conducted to test two hypotheses. First, focusing on behavioral data, we had hypothesized that CRT scores would predict performance on conflict but not on nonconflict problems. This hypothesis was confirmed across both experiments. Building on the findings of Toplak et al. (2011) and Pennycook, Cheyne, et al.

(2014), our findings indicate that the tendency to think reflectively is advantageous when working on problems that pit intuitions against statistical norms. However, because our predictors were positively correlated, we also tested their effects on conflict problems individually (Table 3). With respect to Experiment 1, CRT remained the sole predictor of performance on conflict problems. In contrast, in Experiment 2, word span also emerged as a significant predictor of performance on conflict problems. We believe that word span also emerged as a significant predictor because of the limited time window available for making responses in the fMRI scanner. Specifically, the ability to process linguistic content likely contributed to better performance on the task, as exhibited by the fact that word span predicted performance on both conflict and nonconflict problems in Experiment 2 (both of which require processing linguistic content).

Second, we also found support for the hypothesis that CRT scores would be related to brain activation to a greater extent during conflict than nonconflict problems. Indeed, this was shown to be the case both when we compared brain activation for conflict versus nonconflict problems between participants with high and low CRT scores, as well as when we examined regions of the brain regions where activation covaried as a function of CRT scores in relation to conflict (compared with nonconflict) problems—both analyses pointing to the PCC (Figure 4). Finally, CRT scores were a significant predictor of activation in PCC in conflict problems versus nonconflict problems.

However, contrary to our expectation, CRT-related activation in the brain was localized in the PCC rather than the pFC or ACC. How might PCC function support the kind of thinking that is captured by the CRT? We know that the PCC is a core hub in the default mode network (DMN; Zabelina & Andrews-Hanna, 2016) and that the DMN typically exhibits deactivations during externally directed or difficult tasks. Two possibilities that must be ruled out are that (a) the observed correlation between CRT scores and activation in DMN (i.e., PCC) in conflict problems is due to reduced deactivation in this region in conflict than nonconflict problems and/or (b) there is less deactivation in DMN (i.e., PCC) in conflict problems in participants higher (compared with lower) in CRT. We can rule out the first possibility by observing the pattern of activations in Figure 3B. Specifically, activation in PCC across all participants does not appear to vary as a function of conflict. To address the second possibility, we focused on parameter estimates extracted using MarsBaR from conflict and nonconflict problems as dependent variables and compared differences in activation levels as a function of CRT grouping (high vs. low). We found that on conflict problems activation in PCC was significantly greater in participants with high compared with low CRT scores, $t(42) = 2.15$, $p < .05$, $d = .65$ (Figure 6). In turn, there was no difference between the two groups on nonconflict problems, $t(42) = .03$, *ns*, $d = .01$. Thus,
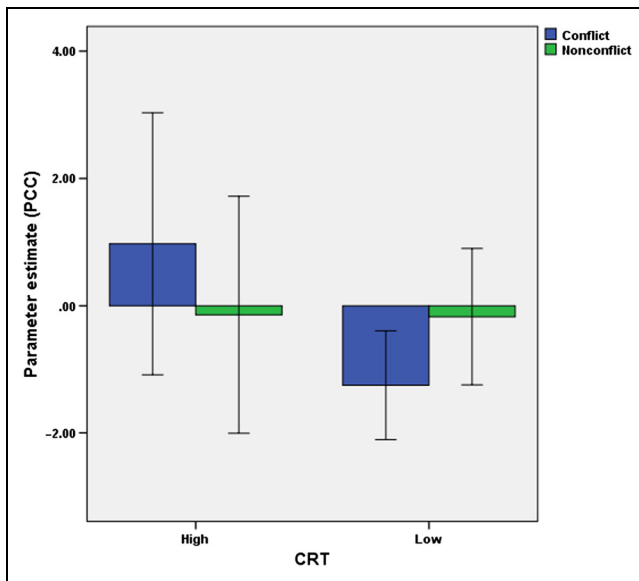
Figure 6. CRT scores and activation in the PCC in relation to conflict and nonconflict problems. Error bars represent 95% confidence intervals.

we can rule out the possibility that there is less deactivation in DMN (i.e., PCC) in conflict problems in participants higher (compared with lower) in CRT. In fact, it appears that there is greater activation in PCC specifically in conflict problems in participants with higher (compared with lower) CRT scores.

In contrast to early conceptualizations that characterized the DMN as a "task-negative network," there is now growing consensus to suggest that it is more optimal to characterize DMN not in terms of its opposition to task but instead "by the self-generated mental content that it supports" (Zabelina & Andrews-Hanna, 2016, p. 87; see also Callard, Smallwood, Golchert, & Margulies, 2013; Andrews-Hanna, Reidler, Sepulcre, Poulin, & Buckner, 2010). In turn, within the DMN and in relation to other large-scale brain networks, PCC appears to play a key role in regulating the focus of attention. For example, Leech and Sharp (2014) recently conducted a large-scale systematic review of the patient and imaging literature on the PCC to argue that not only is there evidence to suggest that the PCC plays a direct role in regulating the focus of attention (Hahn, Ross, & Stein, 2007; Hampson, Driesen, Skudlarski, Gore, & Constable, 2006; Gusnard & Raichle, 2001; see also Pearson, Heilbronner, Barack, Hayden, & Platt, 2011) but that "the role of the PCC could extend beyond supporting internal thought, and rather play a more active role in controlling the balance between an internal and external focus of attention" (p. 23). This conceptualization of the PCC as a regulator of attention between external and internal foci can help us understand its possible functional relationship with individual differences in CRT scores. Specifically, CRT scores could be a reflection of one's ability to switch from intuitive to analytic mode of thought in the service of

cognition (Frederick, 2005). Viewed from this lens, it would appear that the PCC could be involved in facilitating this key feature of reflective thinking measured by the CRT.

Importantly, the ability to switch from an intuitive to an analytic mode of thought presumably also necessitates inhibition; indeed, brain regions that underlie inhibition have been shown to be related to individual differences in CRT performance. Specifically, Oldrati, Patricelli, Colombo, and Antonietti (2016) applied transcranial direct current stimulation to the dorsolateral pFC—a region associated with inhibitory control—to examine whether doing so would modulate performance on the CRT and other mathematical and insights tasks that pitted an incorrect "impulsive" response against a correct deliberative response. Their results demonstrated that, following cathodal stimulations of the dorsolateral pFC (that served to reduce inhibitory control), participants made more errors on the CRT and similar tasks. These findings suggest that the ability to inhibit incorrect prepotent responses is necessary for optimal performance on the CRT. In addition, in conjunction with our results, they suggest that a complete characterization of the neural underpinnings of individual differences involving the CRT will likely involve multiple regions that will contribute various components to this higher-order ability index that measures engagement in reflective thinking.

Interestingly, regarding nonconflict problems, we found that matrix span was the best predictor of performance in Experiment 1, whereas word span emerged as the best predictor of performance in Experiment 2. It is not surprising that, across both experiments, an STM measure would be associated with performance on nonconflict problems, where performance largely depends on the ability to maintain the problem representation in the span of attention without the need to inhibit stereotypical response tendencies or to switch attention to an analytic mode—both of which necessitate the involvement of executive functions. However, we did not expect to find that different measures of STM would emerge as predictors across the two experiments. It is important to note that verbal and matrix spans were strongly correlated in both Experiment 1 and Experiment 2 (Table 3). We believe that the high correlations are consistent with evidence that verbal and visuospatial measures reflect a primarily domain-general construct (e.g., Kane, Hambrick, Tuholski, Payne, & Engle, 2004). As such, the findings across the two experiments regarding nonconflict problems are largely consistent by isolating STM rather than intelligence or CRT scores as the best predictor of performance.

On a more general level, it is also important to take stock of what the joint findings of De Neys et al. (2008) and those reported here indicate about cognitive processes that underlie performance on base rate tasks. First, De Neys et al. (2008) had shown that conflict problems activate ACC more than nonconflict problems and

that ACC was activated regardless of whether participants responded stereotypically or in accordance with base rates. Here we also found that ACC was activated more in conflict than nonconflict problems and that it was also activated when participants responded incorrectly on conflict problems. Together, these findings suggest that, even when participants perform poorly, it is not because they neglect base rates. Activation in ACC suggests that there is some awareness of conflict. Second, we did not replicate De Neys et al.'s (2008) finding that the right IFG was activated more on conflict problems wherein participants had generated base rate rather than stereotypical responses. We argue that one would expect to observe activation in the right IFG if the majority of the participants consistently applied the same strategy (i.e., inhibition) for responding in accordance with base rates throughout the task. However, there is recent evidence to suggest that people can respond in accordance with base rates heuristically (Bago & De Neys, 2017; Newman et al., 2017; Pennycook, Trippas, Handley, & Thompson, 2014). If so, then it is likely that our participants relied on various strategies to arrive at "correct" solutions to conflict problems, which in turn would be associated with a nonhomogenous pattern of neural activation across the sample. Future studies in this area would benefit by manipulating strategy choice through the use of the two-response paradigm (see Newman et al., 2017; Thompson & Johnson, 2014; Pennycook & Thompson, 2012; Thompson, Prowse Turner, & Pennycook, 2011) to examine the neural correlates of correct solution as a function of employed strategies.

We end by highlighting some of the limitations of our experiments that need to be considered while evaluating our findings. First, because of its small sample size, Experiment 1 was likely underpowered for testing hypotheses related to individual differences in performance on heuristics and biases tasks. Second, given the positive correlations reported among intelligence, STM, and CRT scores, both experiments suffered from multicollinearity issues. As such, the relative power of CRT above intelligence and STM for predicting performance on the base rate neglect task and its neural correlates requires replication for assessing its reliability. Third, our neural analyses were based on the prompt time point (see Pennycook, Cheyne, et al., 2014), although the relevant information was available upon the presentation of the base rates. In that sense, the neural correlates of conflict detection can also be examined in relation to that time point. Fourth, our design did not include an independent task that could be used to assess PCC's functional contribution to performance on the base rate neglect task. Certainly, not only do our inferences regarding PCC's function require direct verification, but a full characterization of the neural underpinnings of reflective thinking as measured by the CRT will likely extend beyond PCC. Nevertheless, based on our results, we believe that PCC could play an important role in understanding the neural basis of individual differences in reflective thinking in relation to performance on heuristics and biases tasks.

Notes

1. In two-response paradigms, the participants are instructed to give a first intuitive response as quickly as possible. Afterwards, they get time to reflect and give a second, final response. The two consecutive responses are taken to result from intuitive and reflective processing, respectively.
2. Note that, although responses in line with the base rates are referred to as "correct," strictly speaking the stereotype-based responses do not necessarily represent normative violations. Accuracy here reflects the percentage of responses in line with base rates.
3. Incorrect nonconflict responses were not included in this analysis because of their rarity.
4. In MarsBaR implemented in SPM (Brett et al., 2002), we also produced a spherical ROI (radius = 10 mm) around the PCC found in the first analysis. The activation in the PCC in the second covariation analysis survived cluster-level Bonferroni correction for multiple comparisons within the ROI.

REFERENCES

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron, 65,* 550–562.

Aron, A. R., Fletcher, P. C., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature Neuroscience, 6,* 115–116.

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

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2014). Inhibition and the right inferior frontal cortex: One decade on. *Trends in Cognitive Sciences, 18,* 177–185.

Bago, B., & De Neys, W. (2017). Fast logic?: Examining the time course assumption of dual process theory. *Cognition, 158,* 90–109.

Bissonette, G. B., Powell, E. M., & Roesch, M. R. (2013). Neural structures underlying set-shifting: Roles of medial prefrontal cortex and anterior cingulate cortex. *Behavioral and Brain Research, 250,* 91–101.

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.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences, 12,* 539–546.

Brett, M., Anton, J.-L., Valabregue, R., & Poline, J.-P. (2002). *Region of interest analysis using an SPM toolbox*. Poster presented at the 8th International Conference on Functional Mapping of the Human Brain, June 2–6, 2002, Sendai, Japan.

Callard, F., Smallwood, J., Golchert, J., & Margulies, D. S. (2013). The era of the wandering mind? Twenty-first century research on self-generated mental activity. *Frontiers in Psychology, 4,* Article 891.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research, 169,* 323–338.

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,* 169–187.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE, 6,* e15954.

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

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.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8,* 223–241.

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

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—In search of a phenomenon. *Thinking & Reasoning, 21,* 383–396.

Gusnard, D. A., & Raichle, M. E. (2001). Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience, 2,* 685–694.

Hahn, B., Ross, T. J., & Stein, E. A. (2007). Cingulate activation increases dynamically with response speed under stimulus unpredictability. *Cerebral Cortex, 17,* 1664–1671.

Hampson, M., Driesen, N. R., Skudlarski, P., Gore, J. C., & Constable, R. T. (2006). Brain connectivity related to working memory performance. *Journal of Neuroscience, 26,* 13338–13343.

Harrison, T. L., Shipstead, Z., Hicks, K. L., Hambrick, D. Z., Redick, T. S., & Engle, R. W. (2013). Working memory training may increase working memory capacity but not fluid intelligence. *Psychological Science, 24,* 2409–2419.

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional magnetic resonance imaging*. Sunderland, MA: Sinauer Associates.

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. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

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

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133,* 189–217.

Leech, R., & Sharp, D. J. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain, 137,* 12–32.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41,* 49–100.

Newman, I., 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, & Cognition, 43,* 1154–1170.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9,* 97–113.

Oldrati, V., Patricelli, J., Colombo, B., & Antonietti, A. (2016). The role of dorsolateral prefrontal cortex in inhibition mechanism: A study on cognitive reflection test and similar tasks through neuromodulation. *Neuropsychologia, 91,* 499–508.

Pearson, J. M., Heilbronner, S. R., Barack, D. L., Hayden, B. Y., & Platt, M. L. (2011). Posterior cingulate cortex: Adapting behavior to a changing world. *Trends in Cognitive Sciences, 15,* 143–151.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition, 42,* 1–10.

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,* 34–72.

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.

Prado, J., Kaliuzhna, M., Cheylus, A., & Noveck, I. A. (2008). Overcoming perceptual features in logical reasoning: An event-related potentials study. *Neuropsychologia, 46,* 2629–2637.

Prado, J., & Noveck, I. A. (2007). Overcoming perceptual features in logical reasoning: A parametric functional magnetic resonance imaging study. *Journal of Cognitive Neuroscience, 19,* 642–657.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science, 306,* 443–447.

Robbins, T. W. (1996). Dissociating executive functions of the prefrontal cortex. *Philosophical Transactions of the Royal Society London, Series B, Biological Sciences, 351,* 1643–1670.

Shipley, W. C., Gruber, C. P., Martin, T. A., & Klein, A. M. (2009). *Shipley-2*. Los Angeles: Western Psychological Services.

Simon, G., Lubin, A., Houdé, O., & De Neys, W. (2015). Anterior cingulate cortex and intuitive bias detection during number conservation. *Cognitive Neuroscience, 6,* 158–168.

Slovic, P., & Tversky, A. (1974). Who accepts Savage's axiom? *Behavioral Science, 19,* 368–373.

Stanovich, K. E., & West, R. F. (1998). Individual differences in framing and conjunction effects. *Thinking and Reasoning, 4,* 289–317.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences, 23,* 701–717.

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.

Stupple, E. J., 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.

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, reason, and metacognition. *Cognitive Psychology, 63,* 107–140.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics and biases tasks. *Memory & Cognition, 39,* 1275–1289.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning, 20,* 147–168.

Unsworth, N., & Engle, R. W. (2007). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133,* 1038–1066.

van Veen, V., & Carter, C. S. (2006). Conflict and cognitive control in the brain. *Current Directions in Psychological Science, 15,* 237–240.

Villejoubert, G. (2009). Are representativeness judgments automatic and rapid? The effect of time pressure on the conjunction fallacy. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 30, pp. 2980–2985). Cognitive Science Society.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence (WASI)*. San Antonio, TX: Harcourt Brace, Psychological Corporation.

Zabelina, D. L., & Andrews-Hanna, J. R. (2016). Dynamic network interactions supporting internally-oriented cognition. *Current Opinion in Neurobiology, 40,* 86–93.