Original Article

QJEP

Quarterly Journal of Experimental Psychology
1–21

DOI: 10.1080/17470218.2017.1313283
qjep@sagepub.com
SAGE

Individual differences in conflict detection during reasoning

Darren Frey[1,2], Eric D Johnson[3] and Wim De Neys[1,2,4]

Abstract

Decades of reasoning and decision-making research have established that human judgment is often biased by intuitive heuristics. Recent "error" or bias detection studies have focused on reasoners' abilities to detect whether their heuristic answer conflicts with logical or probabilistic principles. A key open question is whether there are individual differences in this bias detection efficiency. Here we present three studies in which co-registration of different error detection measures (confidence, response time and confidence response time) allowed us to assess bias detection sensitivity at the individual participant level in a range of reasoning tasks. The results indicate that although most individuals show robust bias detection, as indexed by increased latencies and decreased confidence, there is a subgroup of reasoners who consistently fail to do so. We discuss theoretical and practical implications for the field.

Keywords

Reasoning; decision-making; dual-process theory; conflict detection; individual differences

Received: 2 June 2016; revised: 3 March 2017; accepted: 6 March 2017

Suppose you are in charge of hiring for a large firm and you are tasked with finding the next chief technical officer. Although 99% of the applicants whose dossiers you retained come from software engineering backgrounds, a couple are attorneys. On the day of the first interview, you are running late and forget to pick up the candidate's resume from your assistant. He is affable, boisterous and smartly dressed; he makes jokes effortlessly; when presented with a hypothetical conflict to resolve, he talks through both perspectives like a skilled rhetorician. What you have noticed about the candidate neatly coincides with a stereotype, and it is very likely that you intuitively conclude you are interviewing one of the attorneys.

From a logical point of view, coming to this conclusion is questionable. Given that the majority of those whose credentials you retained were engineers, the interviewee is more likely to be an engineer than an attorney; yet, a confluence of conventional associations often leads individuals to neglect a sample's base-rate in this way.[1] Over the course of the past four decades, psychologists, cognitive scientists and behavioral economists have cataloged many such biases, and these have been neatly articulated in the so-called "dual-process framework" (e.g. Kahneman, 2011). Following a dichotomy implicit in the Western tradition at least as early as Aristotle, William James (1890) differentiated between associative thought, which consists of "trains of images suggested one by another" and productive thought, which helps individuals "out of unprecedented situations." Contemporary scholars have elaborated a number of increasingly more refined and detailed dual-process theories. At bottom, most of these contrast bundles of fast, powerful associations (System 1 processes) with slower, more deliberative forms of reasoning (System 2 processes).

Dual-process theories have made convincing claims and have been used to explain how and why reasoners are often biased. The basic idea is that although heuristics generally work quite well and are useful, efficient means of responding to complex environments (Gigerenzer, 2008), they sometimes clash with logical and mathematical principles (Evans, 2003, 2010; Kahneman, 2011; Stanovich & West, 2000). Conflict of this sort is ubiquitous, but until recently, it was generally assumed to be relatively imperceptible.

[1]LaPsyDÉ, UMR 8240, Sorbonne Paris Cité, Paris Descartes University, Paris, France
[2]LaPsyDÉ, UMR 8240, Caen Basse-Normandie University, Paris, France
[3]Department of Basic Psychology and IR3C, University of Barcelona, Barcelona, Spain
[4]LaPsyDÉ, UMR 8240, CNRS, Paris, France

Corresponding author:
Darren Frey, LaPsyDÉ, UMR 8240, Sorbonne Paris Cité, Paris Descartes University, 46 Rue Saint-Jacques, Paris FR-75005, France.
Email: darren.frey@gmail.com

That is, influential authors have suggested that one of the key reasons people end up being biased is precisely that they fail to notice that the heuristic response conflicts with logico-mathematical principles (Evans & Stanovich, 2013; Kahneman, 2011). On this account, although people might have stored the necessary logical principles, they do not activate this knowledge and use it to monitor for potential conflicts with intuitively cued heuristic responses. Consequently, people often end up being biased and are completely oblivious of the erroneous nature of their heuristic response. Yet, a growing body of research suggests that this characterization can be questioned.

Recent studies that focus on empirically examining reasoners' conflict detection efficiency imply that biased reasoners show some sensitivity to conflicts between their heuristic responses and logico-mathematical rules in a range of reasoning tasks (e.g. Bonner & Newell, 2010; De Neys & Glumicic, 2008; Gangemi, Bourgeois-Gironde, & Mancini, 2015; Pennycook, Fugelsang, & Koehler, 2015; Stupple & Ball, 2008; Thompson & Johnson, 2014; for a review, see De Neys, 2014). Recall the opening illustration for a moment; recall the suave, stylized lawyerly interviewee seated before you. The quick, affective associations prompted by his presentation and poise were at odds with the statistical base-rate information that was presented. Regardless of which conclusion an interviewer comes to, she likely feels less confident about it than she would have in the absence of conflicting cues. Exploiting this difference, comparing situations in which there are conflicting cues to those in which there are none (i.e. no-conflict problems), suggests that even biased reasoners do, in fact, tend to detect conflicts. Researchers typically manipulate the congruency of the heuristic and probabilistic, statistical or logical cues to arrive at this conclusion. Returning to the opening example, one needs to simply imagine that 99% of the retained applications were lawyers so that the most statistically likely inference coincides with the heuristic prompted by the stereotype.

There are significant indications that individuals detect conflict across a wide variety of classic reasoning tasks that induce conflict. Behaviorally, when individuals are presented with conflict problems, they respond more slowly (e.g. Bonner & Newell, 2010; De Neys & Glumicic, 2008; Pennycook, Fugelsang, & Koehler, 2012; Stupple & Ball, 2008; Villejoubert, 2009) and they indicate that they are less confident than they are when answering no-conflict problems (De Neys, Cromheeke, & Osman, 2011; De Neys & Feremans, 2013; De Neys, Rossi, & Houdé, 2013; Gangemi et al., 2015; Thompson & Johnson, 2014). Moreover, they tend to make eye movements to the conflicting components of the task while evaluating the conclusion, as evidenced in eye and gaze tracking experiments (Ball, Phillips, Wade, & Quayle, 2006; De Neys & Glumicic, 2008; Morsanyi & Handley, 2012). There is additional neuropsychological evidence for conflict detection from functional magnetic resonance imaging (fMRI; De Neys, Vartanian, & Goel, 2008; Simon, Lubin, Houdé, & Neys, 2015), electroencephalogram (EEG; De Neys, Novitskiy, Ramautar, & Wagemans, 2010) and skin conductance recordings (De Neys, Moyens, & Vansteenwegen, 2010). Although the evidence is taken from a diverse variety of bias tasks and methods, the research is not without its detractors (Aczel, Szollosi, & Bago, 2016; Klauer & Singmann, 2013; Mata, Schubert, & Ferreira, 2014; Pennycook et al., 2012; Singmann, Klauer, & Kellen, 2014; Travers, Rolison, & Feeney, 2016). One of the concerns researchers have voiced is that the findings conflict with traditional characterizations about the laborious and relatively slow nature of logical thinking. Since conflict detection research seems to imply that individuals easily and effortlessly access logical knowledge (see also Handley & Trippas, 2015; Trippas, Handley, Verde, & Morsanyi, 2016 for a related view), these findings are at odds with many popular depictions of logical thinking (Singmann et al., 2014). Even proponents of the successful nature of conflict detection concede that the research is in its formative stages and requires further investigation (De Neys, 2012, 2014).

In this article, we will focus on one important shortcoming of previous conflict detection work. With few exceptions (Mevel et al., 2015; Pennycook et al., 2015; Stupple, Ball, & Ellis, 2013), most previous research in conflict detection deals solely with group-level effects, so the results paint a portrait of the "average" or "typical" biased reasoner. These studies analyze the result for the whole group of biased reasoners averaged across all individuals in the group. In general, such studies allow us to draw conclusions about whether the modal biased reasoner detects conflict, but the potential differences that exist between individuals in the group are largely ignored. In other words, we cannot be sure that all individuals show an effect. This is a critical point since by focusing on the modal reasoner and group-level effects, the initial work on reasoning conflicts might have given readers the erroneous impression that conflict detection is always infallible, an impression no doubt furthered by characterizing conflict detection as "omnipresent" or "flawless" in De Neys et al. (2008) and Franssens and De Neys (2009). Pennycook et al. (2012, 2015) were the first to correct this view and clearly highlight the possibility of conflict detection failures. Here we build on this insight to move beyond group-level findings and categorize primary differences among individual reasoners' abilities to detect conflict. Our goal is to answer a number of key questions: This has left a number of key questions unanswered: Are there individuals who consistently do not detect conflict at all? If so, what portion of the population is this, and how can we account for differences of this sort? So the primary theoretical interest of this analysis is to determine whether and what type of variation exists between individuals in terms of conflict detection, an inquiry others have suggested is

the necessary next step for conflict detection research (De Neys, 2014; De Neys & Bonnefon, 2013; Mata et al., 2014; Pennycook et al., 2015). From a broader perspective, this research is key to further clarifying the relationship between metacognitive processes and reasoning in general (for an overview of the relationship between metacognition and reasoning research, see Ackerman & Thompson, in press).

At a practical level, providing a taxonomy of this sort could help pave the way for more focused debiasing interventions and could have profound educational implications. For example, if it turns out that there are meaningful differences between how individuals detect conflict, then obviously the best strategies for teaching students to combat biases should be tailored to these types (e.g. Lubin, Houdé, & de Neys, 2015; Reyna, Lloyd, & Brainerd, 2003). If individual X is biased because of a conflict detection failure, whereas individual Y shows good detection, then both individuals will likely benefit from different types of training programs that target different components of the reasoning process. Hence, for the optimization of training programs, it is also of paramount importance that we start to characterize the efficiency of conflict detection at the individual level.

To begin to address the question of individual differences in conflict detection sensitivities, we present three studies. In Study 1, we reanalyze data from two previously published experiments. These particular cases were chosen because they used the same measure (confidence ratings) and population (university students) with different tasks (conjunction, base-rate and bat-and-ball tasks—examples of each are found below), and the behavioral data had not been analyzed for evidence of individual differences in conflict detection sensitivities. This reanalysis will allow us to get a first estimate of what portion of the population detects conflict on these tasks and to get a preliminary sense of the variability between subjects.

Study 2 addresses a key methodological problem of individual-level analysis by co-registering three different detection indexes—confidence, response time and confidence response time. Using these three measures sidesteps a primary limitation of our Study 1 and each of the few previous studies that have attempted to look at individual differences in conflict detection: all these studies rely on a single conflict detection measure. Depending on a single measure necessarily complicates interpretation at the individual level. Clearly, one of the key reasons why the initial conflict detection studies (and most of the work in the reasoning and decision-making field) have focused on a group-level approach is that our measures are not perfect. Averaging results across a group of participants allows us to control for the intrinsic measurement noise. When we simply establish whether one individual displays a certain effect (e.g. lower confidence on a conflict vs no-conflict problem), our conclusions become extremely susceptible to such measurement noise. For example, an individual assigning random confidence ratings could end up giving lower confidence ratings on conflict problems compared to no-conflict problems and would be erroneously classified as detecting conflict. By co-registering different measures, we can minimize such misclassification noise. Our estimates of the variability between individuals will be more reliable when they are based on multiple measures that jointly track conflict detection. In other words, if an individual is showing increased latencies in addition to lowered confidence when solving conflict problems, for example, we can be more confident that she is actually detecting conflict and did not simply respond randomly.

Study 2 focuses on one specific reasoning task (i.e. the bat-and-ball problem). Our third study amplifies and extends the analysis by co-registering the three conflict detection indexes on two different reasoning tasks (i.e. conjunction and base-rate task). This enables us to further test the robustness of these measures. In addition, in Study 3, we also start to try to identify one of the cognitive factors that might characterize individuals who do not seem to detect conflict.

Study 1

De Neys et al. (2011) previously demonstrated that despite generally low accuracy rates on bias tasks, erring participants were globally less confident in their choices on conflict items than on control (no-conflict) items in both conjunction and base-rate tasks. De Neys et al. (2013) found similar group-level effects on the notorious bat-and-ball problem. We selected these studies for an individual-level reanalysis because they used the same measure (confidence ratings) and population (university students) with different tasks (conjunction, base-rate and bat-and-ball tasks). Our goal was to get a first estimation of the frequency of (un)successful conflict detection at the individual level. To identify potential individual differences in this reanalysis, we simply contrasted each individual reasoner's confidence on the incorrectly solved conflict and (correctly solved) no-conflict problems. Next, we tallied which percentage of reasoners showed a confidence decrease to get a rough estimate of the percentage of participants who displayed the conflict detection effect (e.g. see Mevel et al., 2015; Pennycook et al., 2015). Before presenting the results, we first give the reader a brief summary of the participants, materials and procedures of the original studies.

Study 1a: base-rate task

Method

Participants. A total of 247 undergraduates at the University of Leuven (Belgium) completed the task in return for course credit (De Neys et al., 2011).

Materials. The participants completed six base-rate problems. Three of these were conflict problems: the description and base-rates cued conflicting responses. The other three were no-conflict problems in which the description and base-rates cued the same response. This is an example of a typical conflict base-rate item:

> In a study 1000 people were tested. Among the participants there were 5 sixteen-year-olds and 995 forty-year-olds. Lisa is a randomly chosen participant of the study.
>
> Lisa likes to listen to techno and electro music. She often wears tight sweaters and jeans. She loves to dance and has a small nose piercing.
>
> What is most likely?
>
> Lisa is sixteen
>
> Lisa is forty

As in the case of the opening example, the narrative description above is designed to cue an intuitive heuristic response based on a stereotype that is at odds with the base-rate. To construct a no-conflict item, De Neys et al. (2011) used congruent stereotypical and base-rate content in this way:

> In a study 1000 people were tested. Among the participants there were 995 dentists and 5 rock singers. Stan is a randomly chosen participant of the study.
>
> Stan is 36. He married his college sweetheart after graduating and has two kids. He doesn't drink or smoke but works long hours.
>
> What is most likely?
>
> Stan is a dentist
>
> Stan is a rock singer.

The stereotyped contents were drawn from a pilot rating study and were verified to moderately but consistently cue one of the two groups in the answer selection. Following each base-rate item, participants were asked to rate their confidence in their previous answers by circling a number on a scale that increased in gradations of 5% from 0% (*not at all confident*) to 100% (*completely confident*).

Note that our primary focus in all reported studies here is biased participants who fail to solve the conflict problem correctly. The results for correct responders are not reanalyzed. First, given the floored accuracy on conflict problems, the group of correct responders is small. Second, given that it is assumed that correct responders also manage to block the heuristic response and thereby resolve the conflict they initially detect, their post-response confidence does not give us a pure indication of conflict detection efficiency per se (i.e. their initial doubt following conflict detection is resolved post-response, e.g., De Neys et al., 2013). This complicates the interpretation of (post-response confidence) conflict detection measures for correct responders. For completeness, the interested reader can find an overview of the correct response data in the Appendix.

Finally, note that as is typically the case in conflict detection studies, as well as in this reanalysis and our additional studies, the rare trials in which no-conflict problems are solved incorrectly are discarded (e.g. De Neys & Glumicic, 2008; Pennycook et al., 2015). Since both the base-rates and description cue the correct response, it is hard to interpret incorrect responses on the no-conflict problems unequivocally.

Procedure. Subjects were administered the experiment at the same time while on a course break. To ensure that the stereotyped contents were balanced, the conflict and no-conflict items were crossed such that half the descriptions used to cue conflicts for half the participants were used in the no-conflict version for the others. Each base-rate item was presented on its own page in a randomized order, except that half of the participants started with a conflict item and the others began with a no-conflict item. The position of the correct response was also randomized.[2]

Results

Group-level findings. For completeness, we first present a summary of the primary group-level findings in the original study and then reanalyze the data for evidence of individual differences. In line with most previous findings, De Neys et al. (2011; Study 1) found that participants were generally biased on the conflict versions of base-rate items, with an average accuracy of 20% (standard error [*SE*] = 1.8). No-conflict items were almost perfectly solved with an average accuracy of 95% (*SE* = 0.83%), $F(1, 246) = 1443.54$, $p < 0.0001$, $\eta^2_p = 0.85$. More crucially, on average, confidence ratings for incorrectly solved conflict problems (68.6%, *SE* = 1.3%) were lower than confidence ratings on correctly solved no-conflict problems (79.0%, *SE* = 1.0%), $F(1, 230) = 59.35$, $p < 0.0001$, $\eta^2_p = 0.21$). Hence, at the group level, biased participants showed a confidence decrease of 10.4% (*SE* = 1.6%) when solving conflict problems. We will refer to this difference as the size of the conflict detection effect. These group-level data are summarized in Table 1 (see "whole biased group").

Individual-level analysis. The group-level results indicate that on average, the group of biased reasoners detects conflict. To identify potential individual differences in this reanalysis, we looked at each individual reasoner's average confidence on the incorrect conflict and (correctly solved) no-conflict problems. Next, we tallied which percentage

Table 1. Overview of individual-level (subgroups of biased reasoners) and group-level (whole biased group) data in Study 1.

Part 1a: base-rate results	Subgroup detection ($n = 153$)	Subgroup reverse detection ($n = 67$)	Subgroup same ($n = 12$)	Whole biased group ($n = 232$)
Proportion of biased reasoners in group (%)	66.0	28.9	5.2	100
Proportion of entire sample (%)	61.9	27.1	4.9	93.9
Average confidence: no conflict correct (%)	82.4 (±1.0)	71.0 (±2.1)	79.3 (±6.0)	79.0 (±1.0)
Average confidence: conflict incorrect (%)	61.7 (±1.5)	82.7 (±1.6)	79.3 (±6.0)	68.6 (±1.3)
Average size of the conflict detection effect (%)	−20.8 (±1.5)	11.7 (±1.0)	–	−10.3 (±1.4)
Detection size–accuracy correlation r (p)	0.30 (0.001)*	0.05 (0.710)	–	0.22 (0.001)*
Part 1b: conjunction results	**Subgroup detection ($n = 61$)**	**Subgroup reverse detection ($n = 35$)**	**Subgroup same ($n = 11$)**	**Whole biased group ($n = 107$)**
Proportion of biased reasoners in subgroup (%)	57.0	32.7	10.3	100
Proportion of entire sample (%)	41.5	23.8	7.5	72.8
Average confidence: no conflict correct (%)	79.3 (±1.7)	58.7 (±2.6)	64.6 (±4.1)	71.1 (±1.6)
Average confidence: conflict incorrect (%)	54.3 (±2.4)	76.7 (±2.1)	64.6 (±4.1)	62.7 (±1.9)
Average size of the conflict detection effect (%)	−25.0 (±1.8)	18.0 (±1.8)	–	−8.4 (±2.3)
Part 1c: bat-and-ball results	**Subgroup detection ($n = 74$)**	**Subgroup reverse detection ($n = 3$)**	**Subgroup same ($n = 119$)**	**Whole biased group ($n = 196$)**
Proportion of biased reasoners in subgroup (%)	37.8	1.5	60.7	100
Proportion of entire sample (%)	29.8	1.2	48.0	79
Average confidence: no conflict correct (%)	96.4 (±1.6)	79.0 (±14.6)	98.8 (±0.6)	97.6 (±0.8)
Average confidence: conflict incorrect (%)	54.5 (±3.7)	89.7 (±9.8)	98.8 (±0.6)	81.9 (±2.1)
Average size of the conflict detection effect (%)	−41.9 (±3.7)	10.7 (±5.2)	–	−15.7 (±2.0)

Effect size reversed for ease of interpretation.
*Significant at $p < 0.05$.

of reasoners showed a confidence decrease when rating incorrect conflict problems as compared to correct no-conflict problems. These results are also shown in Table 1. A total of 232 participants (e.g. 97.9% of the whole sample) gave an incorrect answer on at least one conflict problem. The majority of these biased participants (i.e. 66%) indeed had a marked decrease in their response confidence for the incorrect conflict problems as compared to their rating for correct no-conflict problems. We will refer to this group as the "detection" subgroup. As Table 1 indicates, the average confidence decrease in this group was 20.8% ($SE = 1.5\%$) which is almost twice the size of the decrease that was observed at the group level (i.e. 10.3%, $SE = 1.4\%$). However, there were also 28.9% of biased reasoners who gave a higher confidence response rating on conflict items (mean increase = 11.7%, $SE = 1.0\%$), a subgroup we will call "reverse detection." A further 5.2% of biased reasoners gave the same rating for both types of problems (mean rating = 79.3, $SE = 0.4$), and we will refer to these as the "same" subgroup. Hence, although most biased reasoners showed the effect, there is clearly a smaller subgroup of reasoners who did not show conflict sensitivity as measured by their confidence ratings.

Detection size and conflict accuracy correlations. Previous studies have suggested that the size of the conflict detection effect among biased reasoners (e.g. the size of the confidence decrease when contrasting incorrectly solved conflict and correctly solved no-conflict problems) might be correlated with the response accuracy on the conflict problems (Mevel et al., 2015; Pennycook et al., 2015). The rationale is that the size of the detection effect would reflect the quality or efficiency of the detection process among biased reasoners (Pennycook et al., 2015). Hence, the larger the confidence decrease when one gives an incorrect response, the better one's detection, and the less likely it will be that the individual will err on other conflict problems. In other words, biased individuals with better detection (larger confidence decrease) might be relatively less biased and should obtain higher conflict accuracy scores than biased reasoners with less efficient detection (smaller confidence decrease). We also tested for such a correlation in this reanalysis. Since participants in Study 1a solved a total of three conflict problems, we could calculate for each biased individual who failed to solve at least one conflict problem the total accuracy on the conflict problems.[3] This conflict accuracy score was then correlated with the individual's detection effect size (i.e. the individual's average confidence differences when contrasting incorrectly solved conflict and correctly solved no-conflict problems). The results are also listed in Table 1 (top panel). Note that confidence values were recoded (i.e. we reversed the sign) for this analysis such that

a positive correlation implies that a larger detection effect size (i.e. larger confidence decrease) is associated with higher accuracy. As the table indicates, both for the whole biased group, $r=0.22$, $p<0.001$, and the detection subgroup, $r=0.30$, $p<0.001$, we find a significant correlation between the detection effect size and conflict accuracy. Biased reasoners who show a larger detection effect (i.e. larger confidence decrease) are more likely to solve conflict problems correctly. This supports the claim that the detection effect size reflects the efficiency of the detection process among biased reasoners (Pennycook et al., 2015).

Part 1b: conjunction task

Method

Participants. A total of 147 undergraduates at the University of Leuven participated in the study (De Neys et al., 2011).

Materials. Each participant completed two conjunction problems. For each of the two, participants were first given a short personality description of an individual, modeled on the classic Linda problem. Participants were then provided two statements about the character and asked to indicate which one of the two was most probable. One statement always consisted of a conjunction of two characteristics (one characteristic that was likely given the description and one that was unlikely). The other statement contained only one of these characteristics. Consider the following example of a conflict problem:

> Jon is 32. He is intelligent and punctual but unimaginative and somewhat lifeless. In school, he was strong in mathematics but weak in languages and art.
>
> Which one of the following statements is most likely?
>
> Jon plays in a rock band
>
> Jon plays in a rock band and is an accountant

The conflict above emerges because the cued stereotype features in the conjunctive statement, and the conjunction of any two probabilities is necessarily less likely than either of the conjuncts in isolation (formally: $p(A\&B)\leq p(A)$, $p(B)$). In other words, the likelihood of playing in a rock band is always greater than the likelihood of playing in a rock band *and* being an accountant. To manipulate the conflict nature of the statement, De Neys et al. (2011) changed the content of the non-conjunctive statement. A no-conflict version contains the likely characteristic in the non-conjunctive statement, as is the case below:

> Jon is 32. He is intelligent and punctual but unimaginative and somewhat lifeless. In school, he was strong in mathematics but weak in languages and art.
>
> Which one of the following statements is most likely?
>
> Jon is an accountant
>
> Jon is an accountant and plays in a rock band

People tend to choose the statement that accords with the stereotypical description (Tversky & Kahneman, 1983). They consistently ignore probabilistic considerations and choose the answer that contains content that is most representative of the opening description. Normatively, one should always choose the non-conjunctive statement.[4] So by including the likely content in the non-conjunctive statement, the no-conflict items align normative considerations and intuitive prompts, while the conflict items oppose these by cuing the conjunctive statement.

Procedure. De Neys et al. (2011) presented each participant with one conflict and one no-conflict item. To ensure that the content of the problems was not driving the results, the scenario content and conflict status were crossed. In other words, for half of the participants, the conflict problem was based on the Jon scenario just presented and the no-conflict problem was based on a Linda scenario. The other half of participants saw the opposite: a conflict item based on the Linda scenario and a no-conflict item based on the Jon scenario. As before, the order of the two responses was counterbalanced and each item was presented on its own page of the booklet and followed by the same confidence rating scale as in Part 1. The conflict status and the scenario content of the first problem were counterbalanced. Students completed the tasks at the same time and during a normal break from courses.

Results

Group-level findings. Again, we first present a summary of the group-level findings in the original study and then reanalyze the data for evidence of individual differences. In keeping with the classical findings of Tversky and Kahneman (1983), De Neys et al. (2011; Study 2) found that participants were generally biased on the conflict versions of the conjunction items, with only 24% ($SE=3.5\%$) of respondents solving the conflict item correctly. No-conflict items were consistently solved correctly and had an average accuracy of 96% ($SE=1.6\%$), Wilcoxon matched pairs test, $n=147$, $Z=8.73$, $p<0.0001$. More importantly, on average, confidence ratings for incorrectly solved conflict problems (62.7%, $SE=1.9\%$) were lower than confidence ratings on correctly solved no-conflict problems (71.1%, $SE=1.6\%$), $F(1, 146)=24.49$, $p<0.0001$, $\eta^2_p=0.14$. Hence, at the group level, biased participants showed a confidence decrease of 8.4% ($SE=2.3\%$) when solving conflict problems. These data are also summarized in Table 1 (middle panel).

Individual-level analysis. As was the case in Study 1a, the group-level results suggest that on average, the group of biased reasoners detects conflict. We now turn to the analysis of the individual participants, again considering differences between each individual reasoner's confidence ratings on the incorrectly solved conflict and correctly solved no-conflict problems. These results are also summarized in Table 1. A total of 107 participants (e.g. 72.8% of the whole sample) gave an incorrect answer on the conflict item. The majority of these biased participants (i.e. 57.0%) had diminished confidence in their incorrectly solved conflict problem. As demonstrated in Table 1, the average confidence decrease in this group was 25.0% ($SE = 1.8\%$) which is again more than twice the size of the decrease that was observed at the group level (i.e. 8.4%, $SE = 2.3\%$). There were a number of biased reasoners (i.e. 32.7%) who gave a higher confidence response rating on conflict items (mean increase = 18%, $SE = 1.8\%$) and 10.3% of biased reasoners who gave the same rating for both types of problems (mean rating = 64.5%, $SE = 4.1\%$). Again, the majority of biased reasoners showed the conflict detection effect, but there is still a considerable subset of reasoners who, on this measure, do not seem to be detecting conflict.

Part 1c: bat-and-ball task

Method

Participants. A total of 248 undergraduates at the University of Caen, France, who were in an introductory psychology course participated in the experiment (De Neys et al., 2013).

Materials. Participants were asked to solve the famous bat-and-ball problem (Frederick, 2005) as well as a control version of the same problem. The original (conflict) version of the problem is phrased this way:

> A bat and ball together cost $1.10. The bat costs $1 more than the ball. How much does the ball cost?

The answer that immediately comes to mind is 10 cents, but the correct answer is 5 cents. Although the underlying computation is very simple, most university students (~80%) tend to get it wrong (Bourgeois-Gironde & Vanderhenst, 2009). As Kahneman suggested, participants tend to intuitively substitute the "costs $1 more than" locution with the simpler "costs $1," and so they split the US$1.10 into US$1.00 and US$.10 (De Neys et al., 2013). But of course if one takes a moment to reflect, it is plainly evident that if the ball costs US$1 more than 10 cents, then it costs US$1.10, and the two together would cost US$1.20. De Neys et al. (2013) constructed a no-conflict version of the above that does not lead one astray by replacing the relative phrase ("costs $1 more than") with the statement participants are supposed to be substituting ("costs $1"), thus cuing the correct response. This is an example of a no-conflict version:

> A magazine and a banana together cost $2.90. The magazine costs $2. How much does the banana cost?

Procedure. Participants were presented two items, a conflict and a no-conflict item, both of which were on their own page. Half of the respondents received the conflict version first and vice versa. Item contents and numerical values for both versions of the problem were counterbalanced across participants to ensure that the results were not primarily caused by these constructions, guaranteeing that the results do not emerge because of beliefs about the cost of items or other content features. Following each version of the problem, subjects were asked to write their confidence in their responses on a scale from 0% to 100%.

Results

Group-level results. We again first summarize the primary group-level results found by De Neys et al. (2013). Performance on the conflict bat-and-ball items was relatively low. A mere 21% of individuals solved the latter correctly, which is substantially less than the 98% of individuals who solved the no-conflict version correctly, Wilcoxon matched pairs test, $n = 248$, $Z = 17.6$, $p < 0.0001$.[5] Overall confidence on the incorrectly solved conflict versions (81.9%, $SE = 1.9\%$) was lower than on the correctly solved no-conflict version (97.6%, $SE = 0.6$), $F(1, 94) = 58.54$, $p < 0.0001$, $\eta^2_p = 0.2$. This resulted in group-level confidence decrease of 15.7% ($SE = 2.0\%$).

Individual-level analysis. Again we consider the individual differences between participants' ratings of incorrectly solved conflict items and correctly solved no-conflict items. A total of 196 participants (e.g. 79% of the whole sample) gave an incorrect answer on the conflict item. As shown in Table 1, 74 of these participants (38% of the biased respondents) indicated that they were less confident in the standard conflict version of the bat-and-ball problem. Note that this is considerably less than the 57% and 66% of respondents whose confidence in conflict items diminished in the base-rate and conjunction tasks, respectively. Most of the remaining biased responders showed no difference between the two questions ($n = 119$; 60.7%), and a small number indicated that they were more confident on the conflict items ($n = 3$; 1.5%) As one sees in Table 1, the absolute value of the magnitude of the detection effect of members of the decrease subgroup is more than two and a half times as large as the detection effect of members of the whole group, 41.9% compared to 15.7%. In sum, fewer individuals are showing an effect on this task, as measured by their confidence ratings, but those who do show a very strong effect.

Study 2

Study 1 suggests that there are considerable individual differences in conflict detection. Individual-level analysis reveals that there is a non-negligible group of reasoners on three classical tasks who do not seem to be detecting conflict as measured by their confidence ratings. However, this study has one major shortcoming: the findings rely on one single measure, confidence ratings. The remaining two studies will apply two additional measures that are likely good proxies for conflict detection: response time and confidence response time (Johnson, Tubau, & De Neys, 2016). These additional measures are critical methodological additions. As mentioned earlier, a number of idiosyncratic factors can bias any single measure. Group-level analyses tend to average out this measurement noise, but they can dramatically misrepresent what is happening at the individual level (Baron, 2010). By including additional measures and registering these jointly alongside confidence measures, one strengthens the classification of a particular individual as either detecting or not detecting conflict. If an individual shows consistent detection across different measures, then we can be more certain that the classification is valid. Introducing additional measures, including checks and balances of this sort, is especially important when dealing with individual differences (Stanovich, 2012).

Response time is the amount of time individuals spend on a certain problem, how long they are processing and evaluating their answers; confidence response time is a measure of the amount of time the participants spend assigning their confidence levels. Response time has been linked to uncertainty and conflict detection in a number of studies across a wide variety of tasks (Bonner & Newell, 2010; De Neys & Glumicic, 2008; Pennycook et al., 2012; Scherbaum, Dshemuchadse, Fischer, & Goschke, 2010; Stupple et al., 2013). This research suggests that if one is faced with a decision that causes doubt or uncertainty, apart from having diminished confidence in their responses, reasoners will also take longer to respond than on no-conflict alternatives. Similarly, when faced with a decision about which one is uncertain, evaluating this ambiguity is likely more time-consuming than rendering a verdict about which one is fully confident, so longer confidence response times are also indicative of conflict detection (Johnson et al., 2016; Yeung & Summerfield, 2012).

Method

Participants. Totally, 57 students at the University of Barcelona completed the computer-based experiment in return for course credit.

Materials. The participants were presented the conflict bat-and-ball problem along with a no-conflict version—the same items used in Study 1c (from De Neys et al., 2013)—receiving the conflict item first, while the other half received it last. As in all studies presented here, the content of the problems was balanced. Participants were first presented one of the two problems, which was followed by a text box labeled "cents" where they typed their responses, submitting them by pressing the enter key. The response time measurement is the amount of time that elapses between when the problem is presented and when the participant presses the enter key.

Immediately after submitting their responses to either the conflict or no-conflict items, participants were prompted to enter how confident they were that their answer was correct by typing it on a scale of 0% (*not at all confident*) to 100% (*completely confident*). The measure of confidence latency is the amount of time participants spend generating this confidence assessment, the interval between the presentation of the scale and the moment they press enter.

Procedure. The computer-based test was administered to small groups of students, no more than four at a time. To familiarize participants with the procedure and instructions, they were first presented an unrelated practice math problem, followed by an example confidence query which was displayed on its own page immediately following the problem. After the example and the conflict and no-conflict items, participants were presented with a clearly false statement ("Toulouse is the capital of France") and an additional confidence query as a control to ensure that they were paying attention to these items and not entering confidence responses at random. The average rating on the false statement was 1.58% ($SE = 0.99\%$), with 94.7% of participants entering 0%. Participants were finally asked whether they had previously seen the bat-and-ball problem, all of whom indicated that they had not.

Results

Group-level analysis. As before, we begin by presenting the group-level results. See Table 2 (top panel) for an overview. These are followed by the individual-level analysis. Note that in all analyses using reaction times, we used log-transformed values. However, for the sake of interpretation, the values we report in the tables and the text are raw latencies.

Accuracy. In line with the results of Study 1c and most other studies using the bat-and-ball problem (e.g. Bourgeois-Gironde & Vanderhenst, 2009), most reasoners (82.5%) answered the standard conflict problem incorrectly, so the average accuracy was a mere 17.5% ($SE = 0.05\%$). The no-conflict item was almost perfectly solved with an average accuracy of 98.3% ($SE = 0.02\%$), Wilcoxon matched pairs test, $n = 57$, $Z = 8.7$, $p < 0.0001$.

Table 2. Average group-level findings (*SE*) for each of the three conflict detection indexes as a function of response accuracy in Studies 2 and 3.

Detection index	Conflict: correct	Conflict: incorrect	No-conflict: correct	No-conflict: incorrect
Study 2: bat and ball				
Participants by group	$n=10$	$n=47$	$n=56$	$n=1$
Average confidence (%)	82.8 (±3.7)	84.5 (±3.2)	99.1 (±0.4)	100
Average RT (s)	74.2 (±3.9)	33.6 (±3.8)	14.2 (±3.8)	1.8 (–)
Average confidence RT (s)	5.8 (±0.2)	4.2 (±0.4)	3.5 (±0.4)	2.3 (–)
Study 3: base-rate				
Participants by group	$n=122$	$n=148$	$n=186$	$n=20$
Average confidence (%)	85.1 (±1.5)	81.1 (±1.6)	94.3 (±0.9)	71.3 (±2.8)
Average RT (s)	11.5 (±0.6)	11.8 (±0.5)	9.9 (±0.4)	14.1 (±2.7)
Average confidence RT (s)	3.3 (±0.2)	3.4 (±0.3)	3.9 (±0.2)	3.6 (±0.3)
Study 3: conjunction				
Participants by group	$n=79$	$n=168$	$n=185$	$n=22$
Average confidence (%)	67.1 (±2.9)	73.5 (±.1.3)	86.2 (±1.2)	67.3 (±3.8)
Average RT (s)	8.5 (±0.7)	8.0 (±0.3)	6.5 (±0.3)	7.4 (±1.0)
Average confidence RT (s)	3.1 (±0.1)	3.4 (±0.1)	3.5 (±0.1)	3.3 (±0.3)

RT: response time.

Detection indexes

Confidence. Replicating the earlier findings at the group-level, participants were generally less confident on incorrectly solved conflict items (84.1%, $SE=3.2\%$) than on correctly solved no-conflict items (99.1%, $SE=0.4\%$), amounting to a decrease of 15.0% ($SE=3.5\%$), $F(1, 46)=18.2$, $p<0.001$, $\eta^2p=0.29$. At the group-level, there is, again, evidence of conflict detection using this index.

Response time. As anticipated, and in line with the findings of Johnson et al. (2016), on average participants took longer to respond on incorrectly solved conflict items (33.6 s, $SE=3.8$ s) than on correctly solved no-conflict items (14.2 s, $SE=3.8$ s), $F(1, 46)=36.3$, $p<0.001$, $\eta^2_p=0.45$.

Confidence response time. As Johnson et al. (2016) found, on average, participants did indeed take longer to provide confidence judgments on incorrectly solved conflict items (4.2, $SE=0.4$) than on correctly solved no-conflict items (3.5, $SE=0.4$). However, although there was a trend in the expected direction, the effect did not reach significance, $F(1, 46)=2.3$, $p<0.18$, $\eta^2p=0.05$.

Individual-level analysis. The group-level results replicate previous conflict detection findings with the bat-and-ball problem. To explore the variation between individuals within the sample, we again looked at each participant's individual-level measures and tallied which percentage of biased participants showed the effect. For each individual, we compared the difference between incorrectly solved conflict items and correctly solved no-conflict items for all three indexes. In the present context, longer response times and longer confidence response times on conflict items are the anticipated conflict detection effects, so participants who show these effects are included in the "detection" subgroup in Table 3. As before, such individuals are contrasted with those who show no difference at all between the measures (subgroup "same") and those who have the opposite of the anticipated conflict detection effect (subgroup "reverse detection").

Confidence. At the individual-level, 37% of the participants showed the detection effect and were thus less confident on conflict versions of the problem. This subset of the sample is nearly identical to the proportion of respondents found to have diminished confidence levels in the conflict version in Study 1c (i.e. 38%). As is demonstrated in Table 3 ("confidence effect size"), the average difference between incorrectly solved conflict items and correctly solved no-conflict items of members of this group was 40.5% ($SE=5.4\%$). Replicating our earlier findings, this effect was much more pronounced than the difference found in the whole group (15%, $SE=3.5\%$).

Response time. Although only a minority of individuals seemed to be detecting conflict given the confidence measures, the latencies indicated that most biased participants were indeed sensitive to differences between the conflict and no-conflict items. There was a larger subgroup of individuals (76%) who had increased response times on incorrectly solved conflict versions of the problem compared to no-conflict items. This group took, on average, 28.3 s longer ($SE=4.8$ s) on conflict items, in contrast to the whole group's average of 20.1 ($SE=4.1$).

Confidence response time. A majority of biased individuals also took longer to assign their confidence levels on

Table 3. Individual-level findings for different subgroups of biased reasoners on three conflict detection indexes in Studies 2 and 3.

	Subgroup detection	Subgroup reverse detection	Subgroup same	Whole biased group
Study 2: bat and ball				
Confidence				
% of biased group	37 ($n=17$)	0	63 ($n=29$)	100 ($n=46$)
Confidence effect size (%)	−40.5 (±5.4)	–	0	−15.0 (±3.5)
RT				
% of biased group	76 ($n=35$)	24 ($n=11$)	–	100 ($n=46$)
RT effect size (s)	28.3 (±4.8)	−6.0 (±2.3)	–	20.1 (±4.1)
Confidence RT				
% of biased group	58.7 ($n=27$)	43 ($n=19$)	–	100 ($n=46$)
Confidence RT effect size (s)	2.5 (±0.5)	−2.1 (±0.9)	–	0.5 (±0.6)
Study 3: base-rate				
Confidence				
% of biased group	72 ($n=106$)	16 ($n=24$)	12 ($n=18$)	100 ($n=148$)
Confidence effect size (%)	−20.0 (±1.7)	8.4 (±1.2)	0	−12.3 (±1.6)
Effect size–accuracy correlation r (p)	0.40 (0.001)*	−0.18 (0.39)		0.31 (0.001)*
RT				
% of biased group	64 ($n=94$)	36 ($n=54$)	–	100 ($n=148$)
RT effect size (s)	4.2 (±0.6)	−3.6 (±0.7)	–	1.3 (±0.6)
Effect size–accuracy correlation r (p)	0.25 (0.01)*	−0.21 (0.13)		0.06 (0.44)
Confidence RT				
% of biased group	43 ($n=64$)	57 ($n=84$)	–	100 ($n=148$)
Confidence RT effect size (s)	1.3 (±0.2)	−1.6 (±0.0.4)	–	−0.3 (±0.3)
Effect size–accuracy correlation r (p)	−0.08 (0.53)	−0.08 (0.48)		−0.05 (0.53)
Study 3: conjunction				
Confidence				
% of biased group	79 ($n=132$)	13 ($n=22$)	8 ($n=14$)	100 ($n=168$)
Confidence effect size (%)	−27.6 (±1.1)	10.0 (±2.5)	0	−12.5 (±1.2)
Effect size–accuracy correlation r (p)	0.04 (0.64)	−0.05 (0.84)		−0.001 (0.99)
RT				
% of biased group	71 ($n=120$)	29 ($n=48$)	–	100 ($n=168$)
RT effect size (s)	3.0 (±0.2)	−3.6 (±0.9)		1.2 (±0.4)
Effect size–accuracy correlation r (p)	0.24 (0.01)*	−0.16 (0.28)		−0.05 (0.51)
Confidence RT				
% of biased group	48 ($n=80$)	52 ($n=88$)	–	100 ($n=168$)
Confidence RT effect size (s)	1.1 (±0.2)	−1.3 (±0.2)	–	−0.2 (±0.2)
Effect size–accuracy correlation r (p)	0.20 (0.08)	−0.02 (0.84)		0.09 (0.23)

RT: response time.
* significant at $P < .05$.

the conflict items (59%). Recall that the difference on confidence response time assignments at the group-level was a mere 0.5 s ($SE = 0.6$ s). Among the group of individuals who took longer to determine their confidence levels, the average difference was 2.5 s ($SE = 0.5$ s).

Consistency of detection indexes at individual level across different measures. One could dismiss results of a particular individual detecting (or failing to detect) conflict on *a single measure* as somehow anomalous—a result one might find if participants were assigning confidence ratings at random, for example. However, if people show evidence on more than one measure, then it is more likely that they are, in fact, responding to the conflict. Hence, in a subsequent analysis, for each individual, we simply tallied on how many measures they showed the detection effect, and this result is presented in Table 4.

As the table indicates, the majority of individuals seem to be consistently detecting conflict on at least two different measures. However, although the majority (67%) registers on at least two measures—participants we can classify as *consistent detectors*—the table also demonstrates that a non-negligible proportion of individuals consistently fails to detect conflict across all measures (19.6%) or does so only on a single measure (13.0%). These

Table 4. Proportion of individuals who detect conflict on multiple indexes in Studies 2 and 3.

Number of indexes	0	1	2	3
Study 2: bat and ball	19.6% ($n=9$)	13.0% ($n=6$)	43.5% ($n=20$)	23.9% ($n=11$)
Study 3: base-rate	11.5% ($n=17$)	20.9% ($n=31$)	45.3% ($n=67$)	22.3% ($n=33$)
Study 3: conjunction	3.6% ($n=6$)	20.8% ($n=35$)	50.0% ($n=84$)	25.6% ($n=43$)

findings provide strong evidence for a subset of individuals who tend to not be detecting conflict at all.

Correlations across measures. For our consistency of detection classification, we used a parsimonious, a priori psychometric rationale: showing an effect on more than one measure increases the reliability of the findings. This classification gives us the most assumption-neutral and robust analysis of individual-level data. Nevertheless, one might wonder whether certain measures are more strongly related than others. For example, do reasoners who show consistent detection on two of three measures typically show detection on the confidence and response time indexes or instead on the response time and confidence response time measures? Therefore, it can be informative to look at the actual observed correlations between the detection indexes. Note that confidence values for these correlational analyses were recoded (i.e. we reversed the sign) such that a positive association implies that the expected conflict detection effect was present on both measures. The results showed that there was a strong correlation between the confidence and response time indexes, $r=0.61$, $p<0.001$. Response time also significantly correlated with confidence response time, $r=0.30$, $p<0.04$, but the correlation between confidence and confidence response time did not reach significance, $r=0.12$, $p<0.42$. This suggests that confidence and response time are especially related, with confidence latencies more loosely related. In other words, one may conclude that for those individuals who show detection on two of three measures, these measures are most likely to entail the response time and confidence measures.

Restricted consistency of detection analysis without confidence latency. One may note that the confidence latency index did not reach significance at the group-level analysis and was loosely related to the other indexes in this study. Consequently, one might wonder whether it is warranted to include it in the consistency of detection analysis. Our rationale here was to take a completely a priori and neutral approach for the individual-level analysis. We therefore focused on three behavioral detection measures for which independent previous empirical and theoretical work indicates that they track conflict detection. By restricting the analysis to those measures that show a significant detection effect at the group level in this study, we would make the individual-level classification dependent on the group-level results. Just as a significant group-level effect does not imply that all individuals show the effect, a non-significant group-level effect (or a lack of correlation with other measures) does not imply that the measure is not tracking detection for some individuals. Our independent approach takes the most general and neutral stance in this respect. However, one might nevertheless worry that the inclusion of the confidence latency index is affecting (or potentially biasing) the classification (e.g. artificially boosting the number of consistent detectors). Therefore, we also ran the consistency of detection analysis without the confidence latency index. Key findings are not affected. Obviously, with only two indexes we cannot unequivocally classify participants who only detect on a single index ($n=18$, 39.1% of biased participants). Nevertheless, among those reasoners we could classify, the group of participants who consistently detected on both indexes was still the dominant category ($n=17$, 37.0% of biased participants). A smaller set of individuals failed to detect on both indexes ($n=11$, 24.0% of based participants). Hence, even in this restrictive analysis, consistent detection is more likely at the individual level than consistent non-detection.

Conclusion

In sum, Study 2 indicates that the majority of reasoners seem to be consistently detecting conflict across more than a single measure; however, there is also a subgroup of individuals who consistently fail to do so. This failure to consistently detect conflict across measures suggests that these individuals are genuinely not showing conflict sensitivity on this task.

Study 3

The final study is meant to validate and generalize the findings from Study 2 using the same methods but applying them to different tasks. We opted to focus on the base-rate and conjunction task that were also used in Studies 1a and 1b. Analyzing base-rate and conjunction items in terms of the subjects' response times and confidence response times should broaden the previous findings and expand the taxonomy of individual differences ventured in the first two studies. In addition, Study 3 will also begin to characterize one of the possible cognitive factors that distinguishes those who do not detect conflicts.

Previous research has suggested that one way of accounting for differences in performance on bias tasks is in terms of the reasoner's knowledge stock (Reyna et al., 2003; Stanovich, Toplak, & West, 2008). On a standard bias task that cues an intuitive heuristic response that conflicts with a logical principle, if one simply does not know the relevant principle, then one should naturally fail to generate the correct response. These so-called "mind gaps" (Stanovich et al., 2008) or "storage failures" (De Neys & Bonnefon, 2013) should also lead to a failure to detect conflict. By definition, if one does not know the logical principle, one obviously will not be able to detect that the intuitive response conflicts with it. We can therefore expect that individuals who do not know the relevant reasoning rules will have diminished conflict detection effects as indexed by our three markers, and consequently are likely to be those classified as consistent non-detectors. If right, this helps us isolate one salient factor that can account for individual differences in conflict detection sensitivities.

Method

Participants. There were 195 participants recruited on Amazon's online labor market, Mechanical Turk. The study excluded participants whose IP addresses came from outside of North America.

Materials. Participants answered 16 questions: one block of eight base-rate items and one block of eight conjunction items. Within each block, they were given three conflict, three no-conflict and two control problems. The conflict and no-conflict items were constructed as detailed in Studies 1a and 1b, making use of descriptions that were previously pilot-tested on North Americans (De Neys & Glumicic, 2008; De Neys et al., 2008). The seventh and eighth questions of each block were the control items, which tested for general knowledge of the relevant statistical or probabilistic norms for base-rate and conjunction items without cuing congruent or incongruent stereotypes. These presented a description that was previously judged to be uninformative or neutral with respect to membership of the specified population groups. Examples of the control base-rate and conjunction items are given below.

Base-rate:

> In a study 1000 people were tested. Among the participants there were 995 people who play the trumpet and 5 who play the saxophone.
>
> Tom is 20 years old. He is studying in Washington and has no steady girlfriend. He just bought a second-hand car with his savings.
>
> What is most likely?
>
> Tom plays the trumpet.
>
> Tom plays the saxophone.

Conjunction:

> In a parking lot there are 20 black cars. 15 of the black cars are Volkswagens, 5 of the black cars are Chevrolets. One of the cars in the parking lot has its lights on.
>
> Which one of the following statements is most likely?
>
> The car with its lights on is black.
>
> The car with its lights on is black and is a Volkswagen.

Since the description is neutral, it will not cue a heuristic response. In contrast to the conflict and no-conflict problems, heuristic thinking cannot hinder or aid sound reasoning here. Consequently, solving the neutral problems correctly relies primarily on one's familiarity with the impact of base-rates and conjunctions on probability judgment (De Neys & Feremans, 2013; De Neys & Glumicic, 2008). If one does not know the required logical principle, one should fail to solve the neutral problems (i.e. heuristic thinking cannot help you). If one does know the required logical principle, one should manage to solve the problem correctly (i.e. heuristic thinking cannot bias you). Although participants might still fail to accurately respond for any number of idiosyncratic reasons, these problems can be used as a raw proxy to independently identify and approximate storage failures (De Neys & Feremans, 2013).

To combat ordering effects, half of the participants received the base-rate block first and vice versa, and the position of the normative responses was randomized (i.e. half of the correct answers were presented as option A; half were presented as option B).[6] The order of presentation of the items within the block was randomized, except that the neutral control items were fixed to the two final positions so that they would not prime the participants' subsequent responses. As before, the subjects' response times on the actual problems and on the confidence queries were timed.

Procedure. Although participants were not told they were being timed, they were told that they could take no more than 45 min and they were to complete the questions without interruption. The instruction pages were timed to ensure that the participants read them thoroughly (average time = 16 s; standard deviation [*SD*] = 21 s). Of the 195 original respondents, nine were eliminated because their latencies on one of the measures surpassed a 5 *SD* threshold. The surveys were collected over the course of 25 hr.

Results

Group-level analysis

Accuracy

Base-rate task. Average accuracy on conflict versions (40%, $SE = 1\%$) was substantially lower than on no-conflict versions (96%, $SE = 0.03\%$), $F(1, 185) = 371.4$, $p < 0.001$, $\eta^2_p = 0.67$. Average accuracy on the abstract control problems was 85% ($SE = 0.04\%$).

Conjunction task. The participants' accuracies were similarly distributed on the conjunction items. On average, 22% ($SE = 0.07\%$) of the conflict problems were answered correctly, significantly less than the result on the no-conflict problems (95%, $SE = 0.03\%$), $F(1, 185) = 899.3$, $p < 0.001$, $\eta^2_p = 0.83$. Average accuracy on the control problems was 64% ($SE = 0.06\%$).

Detection indexes. As before, we calculated the group-level differences between indexes on correctly solved no-conflict problems and incorrectly solved conflict problems. We analyzed the data of the two tasks separately. Table 2 (middle and bottom panels) gives an overview of the findings.

Base-rate task

Confidence. Participants were, on average, less confident on incorrectly solved conflict versions of the base-rate problems (81.1%, $SE = 1.6\%$) than on no-conflict versions they solved correctly (94.3%, $SE = 0.9\%$), $F(1, 147) = 69.6$ $p < 0.001$, $\eta^2_p = 0.32$.

Response time. In general, participants tended to take longer on base-rate problems with conflicts that were incorrect (11.8 s; $SE = 0.5$ s) than on those without that were correctly solved (9.9 s; $SE = 50.4$ s), $F(1, 147) = 11.2$, $p < 0.001$, $\eta^2_p = 0.07$.

Confidence response time. The group-level findings on confidence and response times replicate our previous findings. In Study 2, the confidence response time measure had the expected trend but failed to reach significance; however, in this case participants spend, on average, less time rendering confidence judgments on incorrect conflict items (3.4 s; $SE = 0.3$ s) than on no-conflict items (3.9 s; $SE = 0.2$ s), although this effect failed to reach significance, $F(1, 147) = 1.6$, $p < 0.17$, $\eta^2_p = 0.001$.

Conjunction task

Confidence. As was the case on base-rate items, participants were generally less confident on incorrectly solved conflict versions of the conjunction problems (73.5%, $SE = 1.3\%$) than on correctly solved no-conflict versions (86.2%, $SE = 1.2\%$), $F(1, 167) = 107.5$, $p < 0.001$, $\eta^2_p = 0.39$.

Response time. Overall, participants tended to take longer on conjunction problems with conflicts they solved incorrectly (8.0 s; $SE = 0.3$ s) than on those without that they solved correctly (6.5; $SE = 0.3$ s), $F(1, 167) = 30.2$, $p < 0.001$, $\eta^2_p = 0.15$.[7]

Confidence response time. On average, participants tended to take slightly less time to assign their confidence levels on incorrect conflict problems (3.4; $SE = 0.1$) than on correctly solved no-conflict problems (3.5; $SE = 0.1$), $F(1, 167) = 1.0$, $p < 0.31$, $\eta^2_p = 0.006$. As with the base-rate problems, this insignificant trend is the opposite of what was observed in Study 2.

Individual-level analysis.

As in Study 2, for each participant, we calculated the difference between each of the indexes on correctly solved no-conflict problems and on incorrectly solved conflict problems to get a sense of the distribution of individuals registering conflicts on each of the measures, an overview of which is presented in Table 3 (middle and bottom panels).

Base-rate task

Confidence. As before, the majority of biased participants (72%) were less confident in the conflict versions of the problem. The detection effect size—the difference between their confidence measures on conflict and no-conflict items—was −20% ($SE = 1.7\%$), which is nearly twice the effect found at the whole group level (−12.3%, $SE = 1.6\%$). A minority (16%) of biased participants had increased confidence levels on conflict items and 12% had no change at all.

Response time. A majority of respondents (64%) took longer on the incorrectly solved conflict items than on incorrectly solved no-conflict items. On average, they spent 4.2 s ($SE = 0.6$ s) longer on such items, while the entire biased group took a mere 1.3 s ($SE = 0.6$ s).

Confidence response time. In this case, 43% of the participants showed the detection effect, and they spent, on average 1.3 s ($SE = 0.2$ s) longer on conflict than on no-conflict items, with the group average being −0.3 s ($SE = 0.3$ s).

Conjunction task

Confidence. As in Study 1, a large subset of biased participants (79%) were less confident in the conflict versions of the problem. The detection effect size of this group was −27.6% ($SE = 1.1\%$), compared to −18.1% ($SE = 2.2\%$) for the group as a whole. Again, a minority showed the opposite effect (13%) or no difference at all (8%).

Response time. A majority of biased respondents (71%) took longer on the incorrectly solved conflict items

than on incorrectly solved no-conflict items. On average, they spent 3.0 s ($SE = 0.2$ s) longer on such items, while the entire biased group took 1.2 s ($SE = 0.4$ s) longer.

Confidence response time. Nearly half of the biased participants (48%) showed increased confidence response times on incorrectly solved conflict items, spending, on average 1.3 s ($SE = 0.2$ s) longer on these than on no-conflict items. The group as a whole had the opposite effect (−0.2 s; $SE = 0.2$ s).

Detection size and conflict accuracy correlations. In Study 3, participants solved three conflict problems. As in Study 1a, this allows us to examine whether biased reasoners' total accuracy on the conflict problems is correlated with the size of the detection effect. Are individuals with a larger effect relatively less likely to be biased and show higher conflict accuracy? For each of our three detection measures, we calculated the correlation between an individual's detection effect size (i.e. the individual's average difference when contrasting incorrectly solved conflict and correctly solved no-conflict problems) and their total accuracy on the conflict problems. The results are included in Table 3 (middle and bottom panels). Note that confidence values were recoded (i.e. we reversed the sign) for this analysis such that a positive correlation implies that a larger detection effect size (i.e. larger confidence decrease, larger latency increase) is associated with higher accuracy. As Table 3 shows, for the base-rate task, we replicated the association for the confidence measure that was observed in Study 1a: Both for the whole biased group, $r = 0.31$, $p < 0.001$, and for the detection subgroup, $r = 0.40$, $p < 0.001$, we find a significant correlation between the confidence detection effect size and conflict accuracy. However, although there is a significant correlation between the response time effect size and accuracy in the detection subgroup for the base-rate, $r = 0.25$, $p < 0.01$, and conjunction task, $r = 0.24$, $p < 0.01$, the results are less clear for the other measures and conjunction task with typically only small and non-significant correlations. Hence, in these cases there is no clear evidence that the actual size of the detection effect reflects individual differences in the quality of the detection process among biased reasoners. We discuss this further in the general discussion.

Consistent detection: within tasks. Given that results issuing from a single index could occur haphazardly, here we again consider the overall consistency of an individual's responses to conflict on all three detection indexes, the results of which are summarized in Table 4. The general pattern parallels what was found earlier in Study 2, although there are slightly fewer consistent non-detectors (those who registered on none of the three indexes) on these two tasks (11.5% in the base-rate case; 3.6% in the conjunction case compared to 20% for the bat-and-ball task in Study 2). The consistent detectors—those who registered on at least two of the three indexes—accounted for the majority of the biased sample: 67.6% detected consistently in the base-rate case and 75.6% did so in the conjunction case.

Correlations across measures. As in Study 2, one can also calculate correlations between the detection indexes. Note that confidence values for these correlational analyses were recoded (i.e. we reversed the sign) such that a positive association implies that the expected conflict detection effect was present on both measures. For the base-rate task, we observed a correlation between the confidence and response time indexes, $r = 0.27$, $p < 0.001$. Response time also significantly correlated with confidence response time, $r = 0.16$, $p < 0.05$, and there was a marginally significant correlation between confidence and confidence response time $r = 0.14$, $p < 0.08$. For the conjunction task, we observe a strong correlation between the confidence and response time indexes, $r = 0.43$, $p < 0.001$. The correlation between response time and confidence response time, $r = -.07$, $p = 0.36$, and confidence and confidence response time $r = 0.01$, $p = 0.87$, did not reach significance. In sum, in line with the findings in Study 2, these data suggest that confidence and response time show the strongest association. Hence, one can again conclude that for consistent detectors who show detection on two of three measures, these measures are most likely to entail the response time and confidence measures.

Consistent detection: across tasks. The analysis of how consistent individuals are when registering indexes within the two tasks strengthens our claim that at the individual level, most participants are detecting conflicts between intuitively cued heuristic responses and logical or probabilistic norms. As discussed earlier, registering on one single measure on one single task could happen by chance, and the distribution of individuals portrayed from such an anomaly would be consequently biased. So we included additional measures within the task to gain a more accurate classification. A further means of validating our findings is to verify whether participants are consistently detecting conflicts across tasks. One might (rightly) argue that although consistent detection across different measures minimizes the possibility that the classification resulted from mere chance, it does not eliminate it. A reasoner who is assigning confidence ratings at random and by chance erratically attends to the conflict problem might still be erroneously classified as a consistent detector. However, if the consistent detection pattern on task A results from mere chance, it should be extremely unlikely that it is also observed on task B.[8] Hence, by considering how consistently individuals register on each of the indexes across tasks, we further minimize the likelihood of a misclassification.

The overall pattern of detection indexes is depicted in Figure 1, which is a scaled, pictorial cross-tabulation of each of the participants that were biased in both tasks. The results are quite clear, with the majority of individuals (55% of biased reasoners) clustering in the upper right quadrant, where they demonstrated detection sensitivity on two or three indexes on both tasks. The relationship between consistently detecting on one task and on the other is fairly strong. Of those who consistently detected in the base-rate task, 80% did so on the conjunction task ($n = 104$). Of those who consistently detected conflict on the conjunction items, 63% ($n = 134$) did so as well on the base-rate items. The overall correlation between consistent detection across tasks is $r = 0.21$, $p = 0.02$. This provides evidence for the claim that most biased individuals are reliably showing conflict sensitivity in these tasks. At the same time, Figure 1 also indicates that there is a smaller cluster of individuals in the bottom left quadrant (12%) who consistently fail to show consistent detection on more than a single measure in both tasks. This is strong evidence for the existence of a subgroup of participants who reliably fail to demonstrate any conflict sensitivity.

Correlations across tasks. As with the correlations between different detection indexes within a task, one can also compute the correlation between each of the indexes across tasks. For example, do individuals who show a larger detection effect on the base-rate task also show a larger detection effect on the conjunction task? The results of this analysis indicate that the size of the confidence, $r = 0.17$, $p < 0.05$, and response time index, $r = 0.16$, $p < 0.05$, are correlated across tasks. There was no clear association for the confidence response time across tasks, $p = 0.01$, $p = 0.97$. Hence, just as the confidence and response time measures were found to be most strongly associated within the base-rate and conjunction task, these two indexes also show the strongest correlation across both tasks. For completeness, note that a full overview of correlations between the different indexes within and across tasks can be found in Table S5.

Controls. Our results show that although the majority of biased participants detect conflict consistently, there is a smaller subgroup of reasoners for whom this is not the case. One goal of Study 3 was to start exploring what might characterize these individuals. We hypothesized that one factor might be the presence of a "storage failure," and so we added neutral control problems (in which the problem did not cue a heuristic response) as a raw proxy for such a failure. The first thing to recall is that the average scores on the controls were quite high (overall average 75%, $SE = 0.07$; for base-rate items: 85%, $SE = 0.04$; for conjunction items 64%, $SE = 0.06$). This suggests that on average, storage failures are generally quite rare among educated adults. However, whether or not a participant was classified as a consistent detector across tasks indeed tended to be predicted by that participants' total score on the control problems. There was a significant correlation between one's total control accuracy (i.e. the summed base-rate and conjunction control accuracy score, ranging from 0 to 4) and whether or not a participant was classified as a consistent detector, $r = 0.30$, $p < 0.001$. In other words, inconsistent detectors had more evidence of storage failures than consistent detectors as they tended to score lower on our neutral control problems than consistent detectors. This correlational pattern held for the conjunction items alone, $r = 0.17$, $p < 0.025$, and was marginally significant for base-rate problems, $r = 0.14$, $p = 0.09$.

To further illustrate the relation between the control items and conflict detection sensitivities, we also specifically contrasted the average total control score of participants who were classified as consistent detectors across both tasks against others (i.e. participants in the upper right quadrant vs. others in Figure 1). The scores on neutral control items were indeed lower for inconsistent detectors (64% accuracy, average = 2.56, $SE = 0.11$) than for consistent detectors (79% accuracy, average = 3.14, $SE = 0.10$), t-test: $t(117.4) = 12.21$, $p < 0.001$. Again, this pattern was observed both on each of the individual tasks and was significant in the conjunction case, $t(68.98) = 2.16$, $p < 0.035$, and marginally significant in the base-rate case, $t(68.88) = 1.91$, $p = 0.06$.[9]

One might note here that even though performance is lower in the group of those who failed to consistently detect conflict, the absolute performance level is still relatively high. Obviously, storage failure is a sufficient but not necessary condition for failed detection. In other words, although a storage failure is more likely in the inconsistent detector group, most inconsistent detectors do not suffer from a storage failure. Overall, such failures are quite rare. Our point here is simply that when they do occur, they will be most prevalent among the group of inconsistent detectors. To illustrate this further, we also looked at the prevalence of very low control scores (i.e. scores of 0 or 1 out of 4, in other words individuals who failed to solve more than half of the control problems; see Tables S1 and S2 for a complete overview). In the consistent detector group, 2% failed to solve more than half of the control problems; in the "others" group, this figure increased to 11% (and even 17% for the participants in the bottom left quadrant of Figure 1 who consistently failed to detect conflict across tasks). Hence, although low control scores are rare, they are especially concentrated among inconsistent detectors. This implies that although storage failure might not be the most common or important factor that results in a detection failure, its role should not be neglected.

A final statistic that is worth considering is the proportion of inconsistent detectors among those reasoners with a perfect control problem score.[10] For people who show

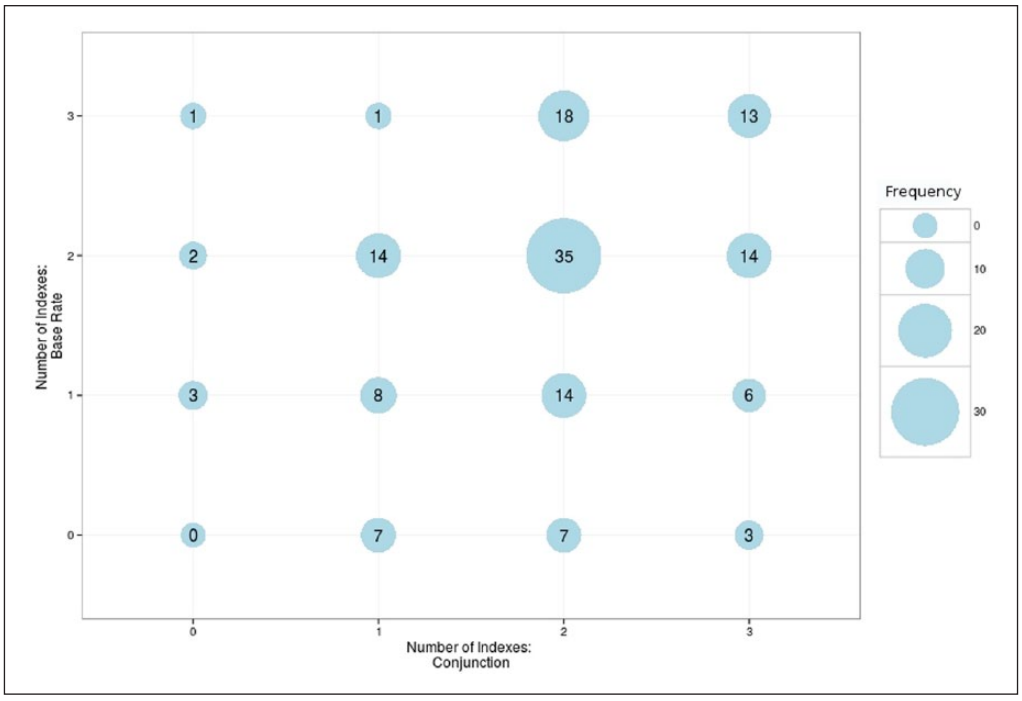

Figure 1. Number of individuals who detect conflict on multiple indexes across the conjunction and base-rate task in Study 3. The majority of participants are clustered in the upper right quadrant in which they consistently detect conflict on at least two measures in both tasks.

perfect control accuracy, we can reasonably eliminate the possibility that they suffer from a storage failure. Hence, a detection failure results from a pure lack of detection per se. This gives us an indication of the prevalence of pure conflict detection failure (i.e. detection failures that cannot be attributed to a storage failure). The results indicate that among the people who showed perfect control accuracy on both tasks (i.e. score of 4 out of 4, $n=48$), 75.0% were classified as consistent detectors and 25.0% ended up in the "other group" (10.4% of the sample were in the more restrictive group of consistent non-detectors; see Table S1). Numbers were comparable when the base-rate (consistent detectors: 71.1%; consistent non-detectors: 12.5%) and conjunction task (consistent detectors: 84.1%; consistent non-detectors: 0%) were considered separately. In sum, in line with the overall pattern illustrated above, the majority of people who have the right mindware also consistently detect conflict across measures and tasks. However, up to 25% of these people will not manage to do so and can be classified as showing evidence of a pure detection failure. The interested reader can find a full split up of the control accuracy by the number of detection indexes in Tables S1 and S2.

Restricted consistency of detection analysis. As in Study 2, we also made sure to run the consistency of detection analysis without the confidence latency index. A full overview of the data can be found in Table S3 and S4 in the Appendix. However, both for the within-task and between-task analyses, key findings were highly similar to the full analysis. As mentioned before, with only two indexes, we cannot unequivocally classify participants who only detect on a single index (base-rate task, $n=54$, 36.5% of biased participants; conjunction tasks, $n=48$, 28.6% of biased participants). Nevertheless, both for the conjunction and base-rate tasks, the group of participants who consistently detected on both indexes was the dominant category (base-rate task, $n=73$, 49.3% of biased participants; conjunction task, $n=102$, 60.7% of biased participants). Only a small minority failed to detect on both indexes (base-rate $n=21$, 14.2% of based participants; conjunction task, $n=18$, 10.7% of based participants). For the analysis across tasks, 31.5% ($n=46$) of participants consistently detected on both measures in both tasks, whereas only 2.7% of participants ($n=4$) failed to detect on both indexes across tasks. Hence, even in this restrictive analysis consistent detection is again far more likely at the individual level than consistent non-detection.

With respect to the control problem analysis, 47.9% ($n=23$) of participants with perfect control accuracy were classified as consistent detectors (detection on both indexes across tasks) and 2.1% ($n=1$) were classified as consistent non-detectors (detection on 0 out of 2 indexes across tasks). These findings were comparable for the individual tasks. In the base-rate case, 52.9% ($n=55$) of participants with perfect control accuracies were consistent detectors and 14.4% ($n=15$) were inconsistent detectors. In the case of conjunction items, 70.7% of those participants with

perfect control accuracy were consistent detectors ($n = 58$), and 7.3% ($n = 6$) were inconsistent detectors. In general, the key pattern of the findings is not affected when the confidence latency index is removed from the analysis.

Discussion

A major shortcoming of previous research on conflict detection during reasoning is that it has predominantly focused on group-level analyses and has largely ignored potential differences that might exist between individuals in the group (De Neys, 2014; De Neys & Bonnefon, 2013; Mata et al., 2014; Pennycook et al., 2015). Here, we reported three studies that help address this issue. In Study 1, we reanalyzed existing group-level data by scoring participants individually for evidence of conflict detection sensitivities on a confidence measure, enabling us to establish a preliminary estimate of the variability among participants. Since single measures are highly susceptible to measurement noise and can lead to the misclassification of participants, in Study 2 we introduced two additional measures, response time and confidence response time, to further assess variation between individuals. In Study 3, we applied all three measures to different tasks and introduced control items to offer a partial account of those who do not detect conflict. Taken together, the results of the three studies indicate that the majority of biased individuals detect conflict on the various tasks that we examined. Studies 2 and 3 clearly indicated that most biased individuals show consistent detection effects across multiple measures and even across tasks. However, at the same time the studies also showed that there is a subgroup of individuals who fail to detect conflict. The size of this subgroup varies based on which measure and task one considers, but in our most conservative estimate in Study 3 (i.e. people who failed to show consistent detection across multiple measures on different tasks) it amounted to about 12% of the group of biased participants. Hence, although conflict detection during reasoning might be successful for most people, this is clearly not the case for everyone. A considerable proportion of educated adult reasoners will not show the detection effects that are observed at the group level.

Study 3 also aimed to identify one of the factors that might characterize those individuals who consistently fail to detect conflict. By definition, if one does not know a certain logical principle, one obviously will not be able to detect that the intuitive response conflicts with it. Consequently, we hypothesized that one factor might be the presence of such a "storage failure." We used neutral control problems (in which the problem did not cue a heuristic response) as a raw storage failure proxy. The results indeed showed that lower scores on the control problems—and hence, a higher likelihood of a storage failure—were more prevalent among the group of inconsistent detectors. However, even among the inconsistent detectors, control performance was high—and storage failures quite rare. This indicates that although our findings show that storage failures are implicated in conflict detection failures (for at least some individuals), they are by no means the most important or most common factor. In other words, most people will fail to detect conflict for reasons other than a storage failure. Therefore, one interesting direction for future studies is to look for other possible moderators that allow us to predict which individuals are most likely to show a conflict detection failure. For example, it might be that reasoners who fail to show consistent detection will be those lowest in numeracy (e.g. Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012; Liberali, Reyna, Furlan, Stein, & Pardo, 2012) cognitive capacity (e.g. Stanovich & West, 2000), showing specific thinking dispositions (e.g. Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014; Stanovich, Toplak, & West, 2008) or a combination of these.

While analysis of individual differences is clearly crucial to conflict detection studies—and, indeed, to judgment and decision-making research more broadly—such work is complicated by questions of the propriety of certain designs and methods (Baron, 2010). As a consequence, we now consider some of our design choices. To begin with, we note that our individual differences' classification was based on simple, a priori defined criteria. We have taken what we understand to be the most neutral and general stance in this respect. We began our individual-level analyses by simply recording whether or not an individual showed an expected detection effect (e.g. Mevel et al., 2015; Pennycook et al., 2015; Travers et al., 2016). We thereby focused on three behavioral detection measures that have been used to track conflict detection at the group level in independent prior studies. Note that we did not restrict the analysis to those measures that showed a significant detection effect at the group level in this study. This would make the individual-level classification dependent on the group-level results. To increase the reliability of our classification, we subsequently looked at the consistency of detection across different measures. One should note that none of these analyses made any further assumptions about the size of the detection effect. For example, any reasoner who shows a confidence decrease and latency increase will be classified as a consistent detector regardless of the size of the latency increase or confidence decrease (e.g. both a 1- and a 10,000-ms increase would equally count as evidence of detection on the latency measure).

There are good reasons for taking this categorical psychometric approach. First, at the theoretical level, it is currently not possible to make a well-informed, justified cutoff value differentiating detection from mere noise (e.g. latency increase needs to be larger than x ms to reflect "genuine" detection). Although there is abundant evidence showing that conflict detection is associated with longer latencies and decreased confidence levels (e.g. Bonner & Newell, 2010; De Neys & Glumicic, 2008; Johnson et al.,

2016; Pennycook et al., 2015; Scherbaum et al., 2010; Stupple et al., 2013; Yeung & Summerfield, 2012), we cannot currently make stronger claims about the precise magnitude of differences that constitutes detection. Second, although it is intuitively appealing to interpret a larger effect as "better" detection, this amounts to a kind of equivocation. The latency increase is a negative byproduct of the experience of conflict. So although a certain amount of slowing down is expected and useful in the presence of conflict, taking too long to reach a decision might in itself be a detriment. It could be that optimal, efficient conflict detection minimizes this negative consequence. In other words, maybe the most efficient conflict detectors are less affected by the conflict and show less processing slow down than less efficient detectors (see Svedholm-Häkkinen, 2015, for a related point). Issues of this sort complicate more graded or qualitative interpretations of the effect sizes. To avoid confusion, note that this argument does not apply to the use of a quantitative interpretation in a purely "statistical sense" (as intended by Pennycook et al., 2015). That is, at least with multiple items, a larger effect indicates that there is more evidence that the person actually detected the conflict. For example, if we assume that actual detection results in a latency increase of size x on an item, on average, people who detect conflict on 75% of the items (i.e. $x \times 0.75$) will have a larger effect than people who only detect conflict on 25% of the items (i.e. $x \times 0.25$). In this sense, a "better" effect does not imply a value judgment about whether the detection is optimal or not but reflects the higher likelihood that there is a true effect.

Nevertheless, previous studies have found interesting relationships between effect sizes and accuracies on conflict items. Our results replicate Mevel et al.'s (2015) work relating accuracy and confidence effect sizes. Specifically, among members who demonstrate the anticipated confidence effect, greater effects are consistently related to higher accuracies. However, we find no such evidence for a relationship between response time measures and accuracy, as was observed in the study of Pennycook et al. (2015). Apart from the concerns about the qualitative interpretation of the magnitude of latencies just mentioned, there are two additional possible explanations for this divergence. First, Pennycook et al. presented many more items than we did (132 vs. 8, including our two controls), which might have amplified a fairly subtle effect. Second, Pennycook et al. used a modified base-rate task that was optimized for response time measurements. This "rapid-response" version, as Pennycook et al. named it, was specifically designed to minimize reading time variances and might well bring out relationships unavailable in the standard version of the base-rate task.

As has been suggested by various scholars, classifying individual differences in conflict detection sensitivities and clarifying the sources of detection failures is a crucial development for the field (De Neys, 2014; Mata et al., 2014; Mevel et al., 2015; Pennycook et al., 2015). Indeed, any theory of reasoning that cannot account for the variability that exists between individuals lacks descriptive precision and predictive power. As previously stressed by Pennycook et al. (2015), the present data highlight that our theoretical models of conflict detection and the nature of heuristic bias need to accommodate the possibility of a conflict detection failure. To illustrate, consider the work by De Neys and Bonnefon (2013) who proposed that classic positions on the nature of bias can be ordered on a timeline from early to late in the reasoning process (see also (Reyna et al., 2003; Stanovich et al., 2008). They argued that the empirical evidence for successful conflict detection supports the idea of biased and unbiased reasoners diverging late in the reasoning process (i.e. after they have both initially accessed stored logical information and detected conflict with a cued heuristic response). The present individual difference findings imply that although this might hold true for most or the modal biased reasoner, there are subgroups of reasoners who will not detect conflict and show an early divergence. This fits with recent claims by Mata et al. (2014) who showed that some individuals are biased because they misrepresent the information in the problem premises before they start the reasoning phase.

Pennycook et al. (2015) recently presented a dual-process model (the three-stage dual-process model) that explicitly encapsulates the possibility of conflict detection failures. As Pennycook et al. clarified, previous work has either capitalized on the success (e.g. De Neys, 2012) or failure (e.g. Evans, 2010; Kahneman, 2011) of conflict detection during thinking. Although few scholars would have argued against possible individuals differences in the detection efficiency, these were not the focus of the research and were not explicitly incorporated into the theoretical models. For example, the De Neys (2012) model does not include or specify a "failed detection" path. Pennycook et al.'s (2015) model includes such a route and is consequently especially well-suited to capture the individual variance that we report here. Although the current findings indicate that the non-detection route is taken by a minority of biased reasoners, it is clear that any viable model needs to include it.

Identifying individual detection variance is theoretically important but might have even further reaching applied implications. Given the importance of sound reasoning for all aspects of life from the classroom to the office, it is not surprising that cognitive and educational scientists have been trying to develop educational "debias" interventions to help people avoid biased thinking. Numerous intervention programs have been developed (e.g. Babai, Shalev, & Stavy, 2015; Evans, Newstead, Allen, & Pollard, 1994; Houdé, 2008; Houdé et al., 2000). However, the results of such interventions have been less than optimal (e.g. Lilienfeld, Ammirati, & Landfield,

2009; Reyna, 2013). We contend that one possible reason is that the programs take a "one-size-fits-all" approach. They are typically targeted at that specific component (e.g. inhibition failure; see Houdé, 2008) that researchers conceive as the modal cause of bias. However, if different individuals are biased for different reasons, they will benefit from a different type of training. Failing to account for such individual differences is bound to limit the efficacy of the intervention. Hence, a straightforward solution to boost the efficiency of intervention programs is to target each type of program at those specific individuals who need them. More specifically, our present findings imply that while training people to inhibit a conflicting heuristic intuition might be a fruitful approach for most reasoners, those individuals with a detection failure will benefit more from a program that helps them monitor for conflict (e.g. Babai et al., 2015). Clearly, if one is not able to detect that a certain intuition is logically inappropriate first, training one's inhibition skills per se will not be effective to remediate the bias. Hence, the individual-level diagnosis that we advocated in this article will also be important in this applied respect. One practical implication that follows from this work is that such individual-level diagnosis can benefit from combining different detection indexes. By administering multiple detection measures, we can boost the reliability of our individual-level classifications.

In closing, we would like to stress that we do not want to argue against the use of a group-level approach or analysis per se. As in many scientific domains, it makes sense to initially start with a group-level exploration of a phenomenon and move to the more complicated individual-level analysis afterward (Stanovich & West, 2000). However, after demonstrating that a certain effect exists, the next step is to determine its prevalence and characterize its natural variability. We believe that the present data present an important step toward this goal and hope that it helps pave the way for a continuation of these individual-level research efforts in conflict detection studies.

Acknowledgements

The authors are grateful to the attendees of these events for their lively discussion and useful feedback. Parts of this article were presented at the 2015 London Reasoning Workshop, the 2015 meeting of the Psychonomics Society (Chicago) and the 2016 International Conference on Thinking.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a research grant (DIAGNOR, ANR-16-CE28-0010-01) from the Agence National de la Recherche. In addition, D.F.'s research is supported by the Sorbonne Paris Cité International Research Grant.

Supplemental material

Supplementary Appendices A1 and A2 and Supplementary Tables S1 to S5 are available at: http://journals.sagepub.com/doi/suppl/10.1080/17470218.2017.1313283.

Notes

1. In what follows, labels such as "appropriate," "correct" or "logical" refer to conclusions that have been deemed normatively correct in classical logic or probability theory, yet the propriety of such notions has been questioned (e.g. Evans & Over, 1996; Stanovich & West, 2000). We adopt the classical characterization for the sake of simplicity throughout. Additionally, we note efforts that have been made to minimize criticisms generated by certain such concerns (see footnote 2).
2. Consistent with the original study, responses that were in line with the base-rates (i.e. selection of the largest group as most likely answer) were labeled as correct answers. Note that using extreme base-rates (995 and 5) and moderate cues, the authors minimized the concern developed by Gigerenzer, Hell, and Blank (1988) that when relying on a formal Bayesian approach, selection of the heuristic response should be considered normatively correct (see De Neys, 2014).
3. Obviously, in studies in which participants solve only one single conflict problem (Study 1b and 1c), this correlation cannot be computed.
4. Again, here we adopt the traditional normative stance, as explained in opening footnote 1. For dissenting views on the conjunction fallacy, see Gigerenzer (1996) and Hertwig and Gigerenzer (1999).
5. For the sake of continuity with the first study, we have included the Wilcoxon statistic, but the original study reports an analysis of variance (ANOVA) $F(1, 247) = 714.94$, $p < 0.0001$, $\eta^2_p = 0.74$.
6. As in all cases throughout, "normative" in this context designates only those choices that are considered normative within traditional logical or classical probability theory.
7. Although these are the results of the log-transformed variables, the raw data manifest the exact same patterns and significance trends.
8. Obviously, this argument only holds if both tasks track the same process. In theory, one's detection efficiency on Task A might be unrelated to detection on Task B. Although the commonality assumption is not unreasonable in the specific case of base-rate and conjunction reasoning tasks, we cannot know this a priori. In case of task specificity, the logic of our argument (detection in Task A can be used to validate detection in Task B) does not hold. However, the observed post hoc association between the detection indexes validates the claim.
9. We use Welch *t*-tests in these cases due to unequal variances.
10. We are indebted to Gordon Pennycook for this suggestion.

References

Ackerman, R., & Thompson, V. (in press). Meta-reasoning: Shedding meta-cognitive light on reasoning research. In L. Ball & V. Thompson (Eds.), *International handbook of thinking & reasoning*. London, England: Psychology Press.

Aczel, B., Szollosi, A., & Bago, B. (2016). Lax monitoring versus logical intuition: The determinants of confidence in conjunction fallacy. *Thinking & Reasoning*, *22*, 99–117.

Babai, R., Shalev, E., & Stavy, R. (2015). A warning intervention improves students' ability to overcome intuitive interference. *ZDM*, *47*, 735–745.

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

Baron, J. (2010). Looking at individual subjects in research on judgment and decision making (or anything). *Acta Psychologica Sinica*, *42*, 88–98.

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.

Bourgeois-Gironde, S., & Vanderhenst, J. B. (2009). How to open the door to System 2: Debiasing the bat and ball problem. In S. Watanabe, A. P. Bloisdell, L. Huber & A. Young (Eds.), *Rational animals, irrational humans* (pp. 235–252). Tokyo, Japan: Keio University Press.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*, 25–47.

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., & Bonnefon, J.-F. (2013). The "whys" and "whens" of individual differences in thinking biases. *Trends in Cognitive Sciences*, *17*, 172–178.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, *6*(1), e15954.

De Neys, W., & Feremans, V. (2013). Development of heuristic bias detection in elementary school. *Developmental Psychology*, *49*, 258–269.

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

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: Autonomic arousal and reasoning conflict. *Cognitive, Affective, & Behavioral Neuroscience*, *10*, 208–216.

De Neys, W., Novitskiy, N., Ramautar, J., & Wagemans, J. (2010). What makes a good reasoner? Brain potentials and heuristic bias susceptibility. In *Proceedings of the Annual Conference of the Cognitive Science Society* (Vol. 32, pp. 1020–1025).

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*, 269–273.

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

Evans, J. S. B. T. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*, 454–459.

Evans, J. S. B. T. (2010). Intuition and reasoning: A dual-process perspective. *Psychological Inquiry*, *21*, 313–326.

Evans, J. S. B. T., Newstead, S. E., Allen, J. L., & Pollard, P. (1994). Debiasing by instruction: The case of belief bias. *European Journal of Cognitive Psychology*, *6*, 263–285.

Evans, J. S. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, UK: Psychology Press.

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

Franssens, S., & Neys, W. D. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*, 105–128.

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.

Gigerenzer, G. (1996). *On narrow norms and vague heuristics: A reply to Kahneman and Tversky*. Retrieved from http://psycnet.apa.org/psycinfo/1996-01780-008

Gigerenzer, G. (2008). *Gut feelings: The intelligence of the unconscious* (Reprint ed.). London, England: Penguin Books.

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base-rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 513–525.

Handley, S. J., & Trippas, D. (2015). Chapter two—Dual processes and the interplay between knowledge and structure: A new parallel processing model. *Psychology of Learning and Motivation*, *62*, 33–58.

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, *12*, 275–305.

Houdé, O. (2008). Pedagogy, not (only) anatomy of reasoning. *Trends in Cognitive Sciences*, *12*, 173–174.

James, W. (1890). *The principles of psychology* (Two volume set 1890). New York, NY: Henry Holt.

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

Klauer, K. C., & Singmann, H. (2013). Does logic feel good? Testing for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1265–1273.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*, 361–381.

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare? *Perspectives on Psychological Science*, *4*, 390–398.

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.

Mata, A., Schubert, A.-L. B., & Ferreira, M. (2014). The role of language comprehension in reasoning: How "good-enough" representations induce biases. *Cognition*, *133*, 457–463.

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*, 227–237.

Morsanyi, K., & Handley, S. (2012). Does thinking make you biased? The case of the engineers and lawyer problem. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *34*, 2049–2054.

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), 1–10.

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

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.

Reyna, V. F. (2013). Psychology: Good and bad news on the adolescent brain. *Nature*, *503*(7474), 48–49.

Reyna, V. F., Lloyd, F. J., & Brainerd, C. J. (2003). Memory, development, and rationality: An integrative theory of judgment and decision making. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 201–245). Cambridge, UK: Cambridge University Press.

Scherbaum, S., Dshemuchadse, M., Fischer, R., & Goschke, T. (2010). How decisions evolve: The temporal dynamics of action selection. *Cognition*, *115*, 407–416.

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

Singmann, H., Klauer, K. C., & Kellen, D. (2014). Intuitive logic revisited: New data and a Bayesian mixed model meta-analysis. *PLoS ONE*, *9*(4), e94223.

Stanovich, K. E. (2012). On the distinction between rationality and intelligence: Implications for understanding individual differences in reasoning. In Holyoak & Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 433–455). New York, NY: Oxford University Press.

Stanovich, K. E., Toplak, M. E., & West, R. F. (2008). The development of rational thought: A taxonomy of heuristics and biases. *Advances in Child Development and Behavior*, *36*, 251–285.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *The Behavioral and Brain Sciences*, *23*, 645–665; discussion 665-726.

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

Stupple, E. J. N., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual-process account from response times and confidence ratings. *Thinking & Reasoning*, *19*, 54–77.

Svedholm-Häkkinen, A. M. (2015). Highly reflective reasoners show no signs of belief inhibition. *Acta Psychologica*, *154*, 69–76.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*, 215–244.

Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109–118.

Trippas, D., Handley, S. J., Verde, M. F., & Morsanyi, K. (2016). Logic brightens my day: Evidence for implicit sensitivity to logical validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 1448–1457.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

Villejoubert, G. (2009). Are representativeness judgments automatic and rapid? The effect of time pressure on the conjunction fallacy. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (*Vol. 30*, pp. 2980–2985). Austin, TX: Cognitive Science Society.

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society, Series B: Biological Sciences*, *367*, 1310–1321.