

PREDICTING INDIVIDUAL DIFFERENCES IN CONFLICT DETECTION AND BIAS SUSCEPTIBILITY DURING REASONING

Jakub Šrol, Wim de Neys

▶ **To cite this version:**

Jakub Šrol, Wim de Neys. PREDICTING INDIVIDUAL DIFFERENCES IN CONFLICT DETECTION AND BIAS SUSCEPTIBILITY DURING REASONING. Thinking and Reasoning, 2021, 27, pp.38-68. 10.1080/13546783.2019.1708793 . hal-03025400

HAL Id: hal-03025400
https://hal.science/hal-03025400

Submitted on 26 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

PREDICTING INDIVIDUAL DIFFERENCES IN CONFLICT DETECTION AND BIAS SUSCEPTIBILITY DURING REASONING

Jakub Šrol[a] & Wim De Neys[b]

[a]Institute of Experimental Psychology, Centre of Social and Psychological Sciences, Slovak Academy of Sciences

[b]Paris Descartes University, LaPsyDE (UMR CNRS 8240), Paris, France

(IN PRESS) - THINKING AND REASONING

Corresponding author:

Jakub Šrol

Institute of Experimental Psychology, Centre of Social and Psychological Sciences, Slovak Academy of Sciences

Dúbravská cesta 9, 841 04 Bratislava, Slovakia

Phone: + 421 - 2 - 5477 5625

Email: jakub.srol@savba.sk

ABSTRACT

One of the key components of the susceptibility to cognitive biases is the ability to monitor for conflict that may arise between intuitively cued "heuristic" answers and logical principles. While there is evidence that people differ in their ability to detect such conflicts, it is not clear which individual factors are driving these differences. In the present large-scale study (N = 399) we explored the role of cognitive ability, thinking dispositions, numeracy, cognitive reflection, and mindware instantiation (i.e. knowledge of logical principles) as potential predictors of individual differences in conflict detection ability and overall accuracy on a battery of reasoning problems. Results showed that mindware instantiation was the single best predictor of both conflict detection efficiency and reasoning accuracy. Cognitive reflection, thinking dispositions, numeracy, and cognitive ability played a significant but smaller role. The full regression model accounted for 40% of the variance in overall reasoning accuracy, but only 7% of the variance in conflict detection efficiency. We discuss the implications of these findings for popular process models of bias susceptibility.

Keywords: conflict detection, bias susceptibility, mindware instantiation, individual differences

PREDICTING INDIVIDUAL DIFFERENCES IN CONFLICT DETECTION AND BIAS SUSCEPTIBILITY DURING REASONING

INTRODUCTION

Several decades of research in the reasoning and decision-making field have shown that even educated reasoners often violate basic logico-mathematical principles (Kahneman, 2011). In general, the problem seems to be that human reasoners have a strong tendency to base their inferences on intuitive rules-of-thumb or "heuristics". Although these intuitive "heuristic" responses will often cue valid problem solutions, they can also conflict with more logical considerations and bias our reasoning. To illustrate, consider the famous bat and ball problem: "A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?" (Frederick, 2005, p.27). Obviously, upon some reflection it is clear that the correct answer is "5 cents" (i.e., 5 cents ball + $1.05 bat = $1.10). However, most educated adults tend to answer that the ball costs 10 cents. The problem seems to be that people intuitively split the $1.10 in $1 and 10 cents and neglect the "more than" statement (De Neys, Rossi, & Houdé, 2013). This intuitively cued "10 cents" answer seems to have an irresistible pull on people's thinking and leads them astray (Kahneman, 2011). And yet, some people are more successful than others at resisting the tendency to go with their heuristic answers when solving such problems, which is at least in part attributable to their ability to detect when their intuition conflicts with the logical consideration of the task at hand (e.g., Frey, Johnson, & De Neys, 2018; Pennycook, Fugelsang, & Koehler, 2015). In the present study, we set out to examine individual difference predictors of this conflict detection ability and their contribution to the overall accuracy on conflict reasoning problems.

Not surprisingly, many theoretical reasoning models have posited that the ability to detect the conflict between intuitively cued heuristic responses and logico-mathematical considerations is critical for sound reasoning (De Neys & Bonnefon, 2013; Evans, 2007; Kahneman, 2011; Stanovich, 2018; Stanovich & West, 2008). Notably, the conflict detection mechanism plays an integral role in traditional default-interventionist models of reasoning (Evans & Stanovich, 2013; Kahneman, 2011; Stanovich & West, 2008). Under such accounts, higher cognitive processes are thought to involve a sequential employment of two types of thought: intuitive (type 1) processing leads to a fast response based on heuristics and initial

problem representation which reasoners can subsequently either affirm or try to correct by engaging in more cognitively demanding analytic (type 2) thought. When dealing with traditional reasoning tasks, such as the bat and ball problem above, the default-interventionist model assumes that intuitive thinking first produces a biased response and thus reasoners need to engage in analytic processing to suppress their intuition and make use of their explicit knowledge to derive the correct answer (Evans & Stanovich, 2013). However, because of the computational costs of analytic thinking, when intuition leads to a response which is not in line with logical considerations of a task at hand, most people will not engage in type 2 processing and thus will not detect this intuition/logic conflict. Hence, according to the traditional default-interventionist model, people are often biased precisely because they fail to detect that their intuitions are in conflict with logical considerations of the task at hand (Kahneman, 2011).

Empirical research on the conflict detection arose precisely to test such predictions derived from the default-interventionist models. In a typical empirical study on conflict detection (for reviews see De Neys, 2012, 2013, 2017), participants are asked to solve reasoning problems from the heuristics and biases literature, as well as their no-conflict counterparts (see Table 1). The two types of tasks are constructed to be as similar as possible in regard to semantic content and their solution requires applying the same logical principles. The only intended difference between them is that while conflict problems are designed to cue heuristic intuitive responses which are in conflict with the solution based on logical norms, in no-conflict tasks intuitive thinking converges with the logical norm in question. The rationale behind conflict detection studies is that if biased reasoners are not detecting the conflict between the logical principle and the intuitively cued response, they would process the conflict problems in the same way as they do the no-conflict ones (De Neys & Glumicic, 2008).

This is, however, not what the available evidence suggests. A large body of conflict detection studies shows that when people give heuristic responses on conflict versions of reasoning problems, they show decreased response confidence in comparison with the no-conflict versions (Frey & De Neys, 2017; Mevel et al., 2014; Stupple, Ball, & Ellis, 2013), prolonged response times (Pennycook et al., 2015; Stupple & Ball, 2008; Swan, Calvillo, & Revlin, 2018), lower feelings of rightness about their answers (Thompson & Johnson, 2014), better recall of information presented in the task (De Neys & Glumicic, 2008), changes in skin conductance (De Neys, Moyens, & Vansteenwegen, 2010), and other neurophysiological

changes (Bago et al., 2018; De Neys, Vartanian, & Goel, 2008; Vartanian et al., 2018). This seems to indicate that even when people are biased, they are at least implicitly sensitive to the fact that their response is not in line with the logically correct response (De Neys, 2012, 2017; however, for critics of this account see Mata, Ferreira, Voss, & Kollei, 2017; Pennycook, Fugelsang, & Koehler, 2012; Singmann, Klauer, & Kellen, 2014).

Table 1. Conflict and no-conflict version of the conjunction fallacy task

Conflict version	No-conflict version
Bill is 34. He is intelligent, punctual but unimaginative and somewhat lifeless. In school, he was strong in mathematics but weak in social studies and humanities.	Bill is 34. He is intelligent, punctual but unimaginative and somewhat lifeless. In school, he was strong in mathematics but weak in social studies and humanities.
Which one of the following statements is most likely?	Which one of the following statements is most likely?
1. Bill plays in a rock band for a hobby	1. Bill is an accountant
2. Bill is an accountant and plays in a rock band for a hobby	2. Bill is an accountant and plays in a rock band for a hobby
	(De Neys & Bonnefon, 2013, p. 175)

Note. In the conflict version of the task, the stereotypical description cues the second option. However, choosing it is considered logically incorrect, as it violates the conjunction rule, i.e. likelihood of two events occurring simultaneously can never exceed the likelihood of one of them occurring separately (Tversky & Kahneman, 1983). In the no-conflict problem, however, both the stereotypical description and the conjunction rule cue the first option.

Given the conflict detection findings, it appeared that people are quickly and effortlessly processing the logical structure of reasoning tasks to detect that their intuition is in conflict with it (De Neys, 2012, 2013, 2017). As these results were not a priori predicted by traditional default-interventionist models (e.g., Evans & Stanovich, 2013; Kahneman, 2011; Stanovich & West, 2008), several authors more recently moved on to advocate so-called *hybrid dual process models* (Bago & De Neys, 2019a; De Neys & Pennycook, 2019; Pennycook, et al., 2015; Thompson & Newman, 2017). While the details of these accounts may differ, they all share the core idea that when people face conflict reasoning tasks, they quickly generate several intuitive responses based on heuristics as well as the logical structure of the problem. Differences in the relative strength of these intuitive outputs determine whether people will detect the conflict and subsequently engage in analytic type 2 processing.

Critically for the present research, early conflict detection studies have typically focused on group-level analyses which indicated that on average people are remarkably good

at detecting conflict (De Neys & Glumicic, 2008; Franssens & De Neys, 2009). This has led researchers to believe that detection failures are unlikely to be a major source of individual differences in accurate reasoning (De Neys & Bonnefon, 2013). Recently, however, the focus has shifted to a more individual–level approach and evidence emerged that people are not at all flawless in their detection ability. Indeed, across several studies, it was observed that at least 10 – 20% of participants did not show any signs of successful detection (Frey et al., 2018; Mevel et al., 2014; Pennycook et al., 2015). Moreover, this figure has been mostly obtained in studies with educated adults, therefore, one might expect that in the general population detection failures could be yet more prevalent. While there now seems ample evidence for substantial individual differences in conflict detection (Frey et al., 2018; Mata et al., 2017; Mevel et al., 2014; Pennycook et al., 2015; Swan et al., 2018), there is very little empirical research available that allows us to identify individual predictors related to the efficiency of this ability.

In theory, a crucial factor for successful conflict detection is the possession of specific mindware necessary to realize that one's intuition is not in line with logical considerations in the task at hand (De Neys & Bonnefon, 2013; Stanovich, 2018). Mindware refers to stored knowledge of elementary mathematical and logical principles necessary for solving the traditional reasoning tasks (Stanovich & West, 2008). Obviously, one will not be able to detect a conflict between an intuitively cued heuristic and a logical principle, if one doesn't know this principle. Recently, Stanovich (2018) has suggested that the degree of mindware instantiation (i.e. the degree to which activation of the principle is automatized) is a key factor in the success of conflict detection. Some preliminary evidence for this claim was already provided by Frey et al. (2018) who included in their research not only conflict and no-conflict tasks (such as those presented in Table 1), but also neutral versions of traditional reasoning problems which served as an indicator of participant's mindware instantiation. Neutral problems were designed to not cue any heuristic responses and thus accuracy on them depended primarily on participant's logico-mathematical knowledge (e.g., the impact of base-rates on probability judgment). Yet, Frey et al. (2018) found mindware instantiation to be only moderately related to the conflict detection ability. This may, however, been in part due to the fact that they used a somewhat limited range of reasoning problems to measure both conflict detection and mindware instantiation (i.e. only base-rate neglect and conjunction fallacy tasks), thus restricting potential correlation between the two variables.

It is also important to note that most of the previous research employed just one type of reasoning problem (e.g., base-rate neglect task), to study conflict detection. This presents a crucial drawback because, as was noted by Frey and De Neys (2017), it is not clear to what extent our detection ability is domain general. That is, we don't know whether people who successfully detect conflict on one problem are also more likely to detect it on other reasoning problems. Indeed, the authors showed that participant's detection ability was not significantly correlated across five different reasoning tasks. A related issue is whether people who show conflict detection as indicated by one index, e.g., decreased confidence, also exhibit detection on other indices. This was examined by Frey et al. (2018), who analyzed whether people consistently detect conflict across three measures: response latency, response confidence, and confidence latency. Their results draw attention to the fact that any single measure is an imperfect indicator of a person's detection ability, and that researchers should simultaneously utilize multiple indices in individual differences studies.

Taking these considerations into account, in the present study we made sure to examine individual differences in conflict detection while employing several indices and reasoning problems to measure participant's detection ability (Frey & De Neys, 2017; Frey et al., 2018). As the available empirical evidence concerning individual variables specifically linked to the conflict detection is sparse, we also decided to examine – in addition to the degree of mindware instantiation – the contribution of a range of common individual difference predictors (i.e., cognitive ability, numeracy, cognitive reflection, and thinking dispositions, see method section for details) as these factors have been hypothesized to be potentially linked to conflict detection ability (Frey et al., 2018; Mevel et al., 2014; Stanovich, 2018). This will allow us to contrast the predictive potential of each individual factor.

In addition to predicting individual differences in conflict detection, the second goal of our study was to examine the role of detection ability and mindware instantiation in participants' overall accuracy on conflict reasoning problems. By employing several reasoning problems to measure participants' susceptibility to cognitive biases as well as a range of standard individual difference predictors, we were able to examine which factors contribute most to the reasoning performance. While individual difference research already established that cognitive ability, numeracy, cognitive reflection, and thinking dispositions all predict conflict reasoning accuracy (e.g., Klaczynski, 2014; Stanovich et al., 2016; Teovanović, Knežević, & Stankov, 2015; Toplak, West, & Stanovich, 2011), to our

knowledge none of the studies have so far investigated these standard individual difference predictors along with indicators of both mindware instantiation and conflict detection ability. This could allow us to critically improve the reasoning accuracy predictions.

METHOD

Participants

Participants were recruited through the Prolific academic online service[1] and paid 7.50£ for their participation. In total, 403 people took part in the study, however, four participants failed to correctly answer two or more of the attention check questions and were excluded from all subsequent analyses. Final sample consisted of 399 participants (32% male, 66% female, 1% other) aged 18 – 73 years (M = 35.81; SD = 12.22). Most participants reported having some college degree (~ 80% of the sample), 15% reported having a high school / GED education, and 2% having less than high school education. The sample size of 400 participants was determined before the data collection began and amounted to the maximum number of participants we could recruit given our research budget. Sensitivity power analysis showed that a sample of this size should provide at least 80% power to detect any correlations of $r > .14$ with 5% error probability.

Materials

Reasoning problems:

Four types of reasoning problems were used in the study. For each type of task, there were four conflict, four no-conflict, and two neutral versions, resulting in 40 items in total. All reasoning problems are included in the supplementary materials. The neutral problems were used to compute the mindware instantiation index (see further).

Syllogistic reasoning task. In syllogisms, participants are presented with two premises and a conclusion and are asked to indicate whether the conclusion follows logically from the

[1] Participants were from United Kingdom, United States, Canada, Australia, or New Zealand, and all reported English as their first language. We do not have information about the distribution of different nationalities in our sample (participants were prescreened about their nationality but not asked about it further), however, most of the participant pool at Prolific academic online services are UK and US nationals (40% and 30%, respectively) at the time we are writing this (15.1.2019).

premises under the assumption that the premises are true. In the *conflict version* of the task, the logical validity of a syllogism is in conflict with the believability of its conclusion (i.e. syllogism is either valid but unbelievable, or invalid but believable). *No-conflict items* were constructed by switching the minor premise and the conclusion and, in the case of unbelievable items, also changing the minor term of the syllogism (see De Neys et al., 2010). This was done to counterbalance problem content across conflict and no-conflict syllogisms. Items were based on the materials in De Neys et al. (2010). Two neutral items were also included which only dealt with abstract statements (i.e. "*All X are Y*"). Internal consistency for the four conflict items was $\alpha = .80$.

Base-rate neglect problems. Every problem provided two types of information about an imaginary person, a proportion of groups in the sample from which the person was randomly drawn (e.g., 5 engineers and 995 lawyers) and a stereotypical description of the individual, which cued one of the groups. Participants were asked to indicate to which of the two groups the imaginary person is more likely to belong. Two versions of the task were created by simply changing the base-rates to favor either the group in line with the stereotypical description (*no-conflict items*) or contrary to the stereotypical information (*conflict problems*). Neutral problems contained a description that did not favor any one of the groups from which the individual was randomly drawn. All items were based on materials used by De Neys and Glumicic (2008). The four conflict items showed good reliability ($\alpha = .82$).

Conjunction fallacy items. Participants were presented with a short stereotypical description of an individual following two statements from which they were supposed to choose the one which was the most likely. One statement always presented a single event pertaining to the described individual (e.g., "*Jake plays the violin*") and the other presented a conjunction of the first event with another feature (e.g., "*Jake plays the violin and is jobless*"). As the probability of the conjunctive statement can never exceed that of a single event, the single event option was always scored as correct. In *no-conflict problems*, this option contained the feature which was also representative of the described individual, while in *conflict problems* the representative feature was part of the conjunctive statement. Neutral problems simply assessed whether participants understood that the probability of a subset of events can never exceed the probability of a superset. Items were based on the material of Frey et al. (2018). Internal consistency of the four conflict items was $\alpha = .78$.

Bat-and-ball problems. Conflict problems were based on the first problem of Frederick's (2005) Cognitive reflection test ("A bat and a ball …") but used different contents and numerical values. *No-conflict versions* were created by eliminating the "more than" statement from the original items (De Neys et al., 2013). In neutral problems, participants simply had to add the values for both items presented within the task. Items were based on the materials from Frey and De Neys (2017). Conflict items showed excellent reliability (α = .94).

Conflict detection indices:

Three measures were used as indicators of participants' conflict detection ability (Frey et al., 2018). First of all, the response latency from the onset of problem presentation until participants submitted a response was recorded. After submitting their response on a reasoning problem, participants were asked to rate their confidence in their answer on a scale of 1 ("*not at all confident*") to 11 ("*absolutely confident*"). The response confidence was used as a second index of conflict detection. Finally, the time that participants took to provide a confidence estimate was recorded and used as a third detection measure. Note that we will also combine the different detection indices in a composite (see results for details) and use it as a predictor of overall conflict reasoning accuracy.

In line with previous work, the conflict detection measures focus on the difference in latency and confidence between incorrectly solved conflict trials and correctly solved no-conflict trials (De Neys et al., 2013; Frey et al., 2018; Mevel et al., 2014; Pennycook et al., 2015). The results for correct responses are not analyzed. Given that it is assumed that in case of correct responding reasoners also managed to block the heuristic response – and thereby resolved the conflict they initially detected – their response latency and confidence does not give us a pure indication of conflict detection efficiency per se (i.e. their initial doubt following conflict detection is also resolved, e.g., De Neys et al., 2013). This complicates the interpretation of conflict detection measures in case of correct responding. Finally, the rare trials in which no-conflict problems are solved incorrectly are discarded (e.g. De Neys & Glumicic, 2008; Pennycook et al., 2015). Since in these problems both intuitive heuristic and logico-mathematical principle cue the correct response, it is hard to interpret incorrect responses on the no-conflict problem unequivocally.

Mindware instantiation:

Neutral versions of reasoning problems were used as a proxy measure of participant's mindware instantiation (Frey et al., 2018). Neutral tasks are similar to the conflict and no-conflict problems but crucially they do not cue heuristic responses. In the absence of a heuristic response which would aid or hinder participant's reasoning, the accuracy on neutral problems depends mainly on the knowledge of logical principles necessary to solve the task at hand. In line with previous studies (e.g., Frey et al., 2018), the average accuracy on individual tasks was very high, as can be seen from Table 2. For all analyses in the present study, we used overall accuracy on all neutral problems combined as an index of participant's mindware instantiation. The eight mindware instantiation items showed relatively poor reliability (α = .28), which was likely caused by very high performance on all neutral reasoning problems, and thus strong ceiling effect on these items. Furthermore, low reliability might also result from the fact that mindware instantiation is quite task specific (Stanovich, 2018).

Note that while mindware instantiation is used as a predictor throughout the analyses in the present study, it will be treated separately from the standard individual difference predictors due to its hypothesized distinct theoretical role in conflict detection and overall reasoning accuracy (De Neys & Bonnefon, 2013; Stanovich, 2018).

Table 2. Accuracy on neutral reasoning problems

	M	*SD*
Syllogistic reasoning tasks	71%	28.93
Base-rate neglect problems	84%	28.68
Conjunction fallacy items	73%	38.43
Bat-and-ball problems	99.8%	3.54
Overall performance	82%	14.99

Note. The table contains mean accuracies (in %) and their standard deviations for the neutral versions of reasoning problems

Standard individual difference predictors:

Cognitive ability. To measure people's cognitive ability we used the *Vienna matrix test* (VMT; Klose, Černochová, & Král, 2002). VMT is a standardized cognitive ability test which resembles the Raven's progressive matrices. The Czech adaptation of the VMT that was used in the present study shows a correlation of r = .92 with Raven' test (Klose et al., 2002). It originally consists of 24 items of increasing difficulty in which participants need to find a pattern in a complex 3 x 3 picture matrix and choose one of the eight options to complete it under a 24-minute time limit. To reduce participants load due to the study length, we decided to adopt a shortened, 14-item version with a 15-minute time limit based on the

data collected in a previous study. Šrol (2019) showed that this shortened version has good reliability (α = .82) and a very high correlation with the full version of the test (r = .96).

Thinking disposition measures. Thinking dispositions are related to people's epistemic values and self-regulation and entail propensities for different types of thought, such as the tendency to consider opposing views before reaching any conclusions, or to think extensively about a problem before responding (Stanovich, West, & Toplak, 2016). To tap participants' analytic thinking disposition, a short 5–item *Need for Cognition* scale (NFC; example item: "*I prefer complex to simple problems*") was used. Similarly, we employed a 5–item *Faith in Intuition* scale (FI; example item: "*I believe in trusting my hunches*") to measure participants' inclination toward intuitive thinking. Both scales were taken from the work of Epstein, Pacini, Denes-Raj, & Heier (1996). Epstein et al. report high correlations between both 5–item versions and their original longer counterparts (r = 90. for NFC; r = .85 for FI). In the present study, participants were asked to rate the items of both tests on a scale from 1 ("completely uncharacteristic of me") to 5 ("completely characteristic of me"). The NFC and FI were intended to reflect two independent processing styles rather than opposite ends of a single dimension (Epstein et al., 1996) and in line with this they tend to be uncorrelated (in the present study: $r(399) = –.09$, p = .072). Therefore, in all analyses here we treated them as two separate individual difference predictors rather than a single thinking disposition composite.

Numeracy. Two methods were used to measure participants' numeracy. The first one was the *Berlin numeracy test* (BNT; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012), a four-item measure which in an extensive validation study showed good discrimination and convergent validity with other measures of numerical ability in diverse samples. As the open-ended version of the test is usually quite hard for use in the general population, a multiple-choice format of the BNT was used here (see Appendix in Cokely et al., 2012). Despite the multiple-choice format, the four-item test showed very low reliability in the present study (α = .41), presumably because most items showed up to be too hard for our participants (M = 1.62, SD = 1.08). Second, we also included the self-report *Subjective numeracy scale* (SNS; Fagerlin et al., 2007). The SNS consists of 8 questions pertaining to the ability to use numerical information (e.g., "*How good are you at working with fractions?*") and preference for numerical over other formats of information (e.g., "*How often do you find numerical information to be useful?*") to which participants respond using a 6-point scale. In the present study, the scale exhibited very good reliability (α = .86). The average rating on the eight items was 3.88 (SD = 1.05). Because of the low reliability of BNT and moderate

correlation between the two numeracy methods employed in the present study (r = .25)[2], we decided to normalize the scores on all 12 items of the two numeracy measures combined and compute the average of the normalized scores to create a single numeracy composite. This composite value is used as an index of the participant's numeracy throughout the study[3].

Cognitive reflection measure (CR). The cognitive reflection measure was modeled after Frederick's (2005) Cognitive Reflection Test. Four items were taken from Thomson and Oppenheimer (2016). Six additional items were based on Šrol (2019). The scores on the four and six item test were highly correlated (r = .57) and were summed to form a single composite. Note that the bat-and-ball problem was not among the items.

The descriptive statistics for the standard individual difference measures are reported in Table 3.

Table 3. Descriptives of standard individual difference predictors

	M	*SD*	*α*
Cognitive ability – Vienna matrix test	5.31	3.11	.75
Thinking dispositions – Need for Cognition	3.50	0.77	.79
Thinking dispositions – Faith in Intuition	3.65	0.76	.87
Numeracy	0.00	0.56	.81
Cognitive reflection	5.21	2.45	.75

Note. The table contains mean scores, standard deviations, and reliability estimates (Cronbach's α) for the standard individual difference predictors employed in the study.

Procedure

The study was created with the Qualtrics software package and run online. It consisted of two blocks of materials, one containing the reasoning problems and the other containing the individual difference measures and one additional measure which is not reported in the present study. The order of the blocks was randomized between participants and the order of

[2] The low correlation between Berlin numeracy test scores and Subjective numeracy scale here is likely influenced by low reliability of the former method. For example, Fagerlin et al. (2007) report much stronger relationship between SNS and a different performance-based measure of numeracy (r = .53).

[3] To ensure that combining an objective and subjective numeracy measure into a single composite did not confound our results, we have also run all of the of analyses pertaining to individual difference predictors of conflict detection and conflict problem accuracy while including the two numeracy measures as separate variables. The analyses are available in Section G of the supplementary material. Note that the results are completely consistent with the key conclusions presented in the main manuscript.

materials within each block was randomized within participants. Before every type of reasoning problem, participants were presented with instructions and an example item to familiarize them with the tasks. All reasoning problems except for bat-and-ball items were presented in two steps in order to reduce variability in response latencies due to reading (e.g., Frey et al., 2018). Participants first saw only the problem description, i.e. the first two premises of syllogisms, base-rate information in the base-rate neglect task, and the description of an individual in conjunction fallacy task. Then, the actual question was presented – the conclusion of a syllogism, description of an individual in base-rate neglect task, the two possible statements about the individual in the conjunction fallacy task – along with two response choices. Participants responded by selecting one of the response choices and submitted their response by clicking on a button to move to the next page. The time from the onset of the actual question presentation until they submitted their response was recorded. Participants were not explicitly told their responses to reasoning problems were timed but were instructed not to take breaks while solving these problems and to submit their answers immediately after deciding on the response. Three attention check questions were included in the study and participants who answered less than two of them correctly were automatically dropped from further analyses. Two attention check questions were mixed with base-rate neglect items and the cognitive reflection measure and were created to resemble these materials but to have an unambiguous correct response. One item was included in the BNT where participants were explicitly asked to always choose the option ”*none of the above*” to indicate they had read the item.

RESULTS

The main aim of the present study was to identify individual predictors of reasoners’ conflict detection ability and overall reasoning accuracy. However, for consistency with previous research, we first present the results of traditional conflict detection analyses where we examine the differences in participant’s response latency, confidence, and confidence latency between conflict and no-conflict reasoning problems both in the entire sample (group-level analysis) and individually for participants who showed signs of successful conflict detection (individual-level analysis). We then explore the generality of the conflict detection ability by looking at the correlations between the ability to detect conflict across different detection indices and reasoning problems. Next, we conduct a correlation analysis to establish the

mutual relationships between conflict detection efficiency, overall reasoning accuracy, mindware instantiation, and our standard individual difference predictors. Finally, we present two regression analyses in which we examine relative contributions of standard individual difference predictors and mindware instantiation to conflict detection efficiency and overall reasoning accuracy.

Group–level reasoning accuracy and conflict detection analyses

Table 4 shows an overview of the group-level reasoning accuracy and conflict detection findings. All results were analyzed separately for the four reasoning problems but for simplicity, we also calculated the overall performance across all conflict and no-conflict problems. Consistently with much previous research, overall accuracy for no-conflict problems (M = 93%, SD = 8.62) was much higher than for their conflict (M = 43%, SD = 28.42) counterparts, $t(398) = 35.54$; $p < .001$; $d = 1.78$. As Table 4 shows, this pattern was observed on every individual reasoning task.

More importantly for the present study, for every participant we computed the difference in average response latency[4], confidence, and confidence latency for incorrectly solved conflict and correctly answered no-conflict problems (e.g., Frey et al., 2018). Participants who did not give any incorrect answers on conflict or correct answers on the no-conflict items were dropped from the respective analyses (n's are indicated in Table 4). Averaged across all reasoning problems, participants took longer to answer conflict problems than the no-conflict ones, $t(384) = 11.64$; $p < .001$; $d = 0.59$; and the former were associated with lower response confidence that the latter, $t(384) = 14.17$; $p < .001$; $d = 0.72$. By and large, the overall results were observed on each separate task. While there was no difference in overall confidence latency observed on conflict and no-conflict problems, $t(384) = 0.06$; $p = .95$; $d = 0.00$; participants took significantly longer to provide confidence estimates for conflict syllogisms and bat-and-ball tasks. As is evident from Table 4, participants showed classic conflict detection signs in most of the individual reasoning problems and detection

[4] Prior to the analyses, all latency data were checked for outlying observations. Latency values which were more than three standard deviations above/below the mean of the respective index were replaced with the value of three standard deviations above/below the average. All analyses reported in the main manuscript and the supplementary material are based on the outlier treated data. However, all analyses were also run on the raw data (before outlier replacement) and the results were consistent with the conclusions presented in the study.

indices. Exceptions were response confidence in the syllogistic reasoning task and confidence latency in base-rate and conjunction fallacy problems.

Table 4. Summary of group–level reasoning accuracy and conflict detection analyses

	Accuracy	Response latency	Response confidence	Confidence latency
Syllogistic reasoning task				
no-conflict (*SD*)	86% (18.86)	4.78 (2.84)	9.98 (1.43)	2.46 (1.16)
conflict (*SD*)	55% (38.74)	5.52 (4.20)	9.67 (1.65)	2.66 (1.40)
difference (*n* = 270)	$t(398) = 16.60$ ***	$t(269) = 3.30$ **	$t(269) = 1.43$	$t(269) = 2.20$ *
	$d = 0.83$	$d = 0.20$	$d = 0.09$	$d = 0.13$
Bat-and-ball items				
no-conflict (*SD*)	98% (10.51)	8.48 (4.10)	10.65 (0.84)	2.31 (1.36)
conflict (*SD*)	42% (45.68)	12.21 (8.92)	10.05 (1.92)	2.61 (1.68)
difference (*n* = 273)	$t(398) = 22.67$ ***	$t(272) = 8.30$ ***	$t(272) = 5.71$ ***	$t(272) = 2.61$ *
	$d = 1.13$	$d = 0.50$	$d = 0.35$	$d = 0.16$
Base-rate neglect problems				
no-conflict (*SD*)	94% (13.16)	10.87 (5.75)	8.96 (1.58)	2.67 (1.20)
conflict (*SD*)	47% (39.95)	12.70 (7.99)	7.78 (1.94)	2.65 (1.21)
difference (*n* = 294)	$t(398) = 23.57$ ***	$t(293) = 5.03$ ***	$t(293) = 9.47$ ***	$t(293) = 0.33$
	$d = 1.18$	$d = 0.29$	$d = 0.55$	$d = 0.02$
Conjunction fallacy problems				
no-conflict (*SD*)	93% (16.85)	7.89 (4.15)	7.96 (2.10)	2.79 (1.20)
conflict (*SD*)	30% (35.06)	10.80 (6.11)	6.79 (2.17)	2.70 (1.08)
difference (*n* = 341)	$t(398) = 35.50$ ***	$t(340) = 10.56$ ***	$t(340) = 13.92$ ***	$t(340) = 1.79$
	$d = 1.68$	$d = 0.57$	$d = 0.75$	$d = 0.10$
Overall performance				
no-conflict (*SD*)	93% (8.62)	7.92 (3.25)	9.39 (1.06)	2.66 (3.12)
conflict (*SD*)	43% (28.42)	10.49 (5.54)	8.28 (1.71)	2.65 (1.11)
difference (*n* = 385)	$t(398) = 35.54$ ***	$t(384) = 11.64$ ***	$t(384) = 14.17$ ***	$t(384) = 0.06$
	$d = 1.78$	$d = 0.59$	$d = 0.72$	$d = 0.00$

Note. Overall performance reflects participants' mean accuracy, response latency, confidence, and confidence latency averaged across all four reasoning tasks. Response latency data are reported in seconds. Cohen's *d* is reported as a measure of effect size. * $p < .05$; ** $p < .01$; *** $p < .001$

Individual-level conflict detection analyses

Following Frey et al. (2018), for every detection index, participants who got at least one conflict item incorrect and one no-conflict item correct (whole biased group) were further divided into three subgroups according to whether they showed longer latencies and lower confidence for incorrect conflict than correct no-conflict problems (*detection subgroup*), the opposite pattern of results (*reverse detection*), or the same latency and confidence estimates for the two versions of problems (*same subgroup*). Here we only present the overview of results for the detection subgroup but a complete summary of individual-level conflict detection analyses can be found in the supplementary material. Table 5 presents the proportions of participants who successfully detected conflict, as well as the detection effects (i.e., the average difference in response latency, confidence, and confidence latency between

conflict and no-conflict problems) across the reasoning problems and detection indices. Results were analyzed separately for the four reasoning problems but we again also calculated overall performance across all conflict and no-conflict problems for simplicity.

Considering the overall performance on all conflict and no-conflict reasoning problems combined, across the three indices 54 – 81% of biased reasoners showed signs of successful conflict detection. This high prevalence of successful detection was also observed in most of the individual reasoning tasks. Two exceptions were found in case of the response confidence index in syllogistic reasoning and bat-and-ball problems, where the proportion of participants exhibiting successful detection was somewhat lower. In sum, consistent with other studies that employed individual-level conflict detection analyses (Frey et al., 2018; Mevel et al., 2014; Pennycook et al., 2015) our results show that while on average most reasoners may be quite capable of detecting the misleading nature of their intuitions, substantial individual differences can nevertheless be observed on particular reasoning problems and detection indices.

Table 5. Summary of individual-level conflict detection analysis for the detection subgroup

	Detection index		
	Response latency	Response confidence	Confidence latency
Syllogistic reasoning task			
proportion of biased group (*n* = 270)	153 (57%)	97 (36%)	146 (54%)
conflict detection effect (*SD*)	–2.75 (3.38)	–1.50 (1.33)	–0.99 (1.44)
Bat-and-ball items			
proportion of biased group (*n* = 273)	208 (76%)	79 (29%)	152 (56%)
conflict detection effect (*SD*)	–5.78 (7.11)	–2.16 (2.43)	–1.13 (1.77)
Base-rate neglect problems			
proportion of biased group (*n* = 294)	169 (57%)	188 (64%)	140 (48%)
conflict detection effect (*SD*)	–4.97 (6.25)	–2.05 (1.60)	–0.72 (0.85)
Conjunction fallacy problems			
proportion of biased group (*n* = 341)	268 (79%)	257 (75%)	168 (49%)
conflict detection effect (*SD*)	–4.44 (4.37)	–1.67 (1.29)	–0.59 (0.60)
Overall performance			
proportion of biased group (*n* = 384)	311 (81%)	294 (77%)	208 (54%)
conflict detection effect (*SD*)	–3.41 (3.82)	–1.22 (1.13)	–0.64 (0.98)

Note. Response latency data are reported in seconds.

Correlations of detection efficiency across detection indices & reasoning problems

To find out whether people consistently detected conflict across the three detection indices, we have computed their detection efficiency separately for response latency, response

confidence, and confidence latency measures. The detection efficiency was calculated for every participant as the number of times they showed a successful detection on a given index divided by the total number of reasoning tasks on which they were biased (Frey & De Neys, 2017)[5]. We calculated the amount of successful detections by summing the number of times participants showed either lower confidence, longer response latency, or longer confidence latency on the conflict in comparison with no-conflict versions of the four reasoning problems. As successful detection is calculated only from reasoning problems on which participants are biased and respond incorrectly, those problems on which a participant did not give any incorrect conflict responses were not used to calculate their detection efficiency. Therefore, we divided the amount of successful detections for every given participant by the total number of times they could have detected the conflict on the four reasoning problems (i.e. the number of times they answered incorrectly). Participants who did not give any incorrect responses to conflict problems (n = 15) were dropped from subsequent analyses.

The detection efficiency index calculated on the basis of response latencies was correlated with the one based on confidence, $r(384) = .238$, $p < .001$, but not with the one based on confidence latencies, $r(384) = .067$, $p = .19$. Confidence and confidence latency efficiencies were weakly correlated, $r(384) = .116$, $p = .02$. Given that confidence latency detection efficiency was at best weakly related to the other indices and previous research already questioned the reliability of this index (Frey et al., 2018), we decided to drop confidence latency from all subsequent analyses and focus on the two remaining indices (confidence and response latency) in the rest of the results. For completeness, the analyses pertaining to the confidence latency index can be found in the supplementary material.

While the abovementioned results show some consistency in successful detection based on response latency and confidence detection indices, we were also interested in whether participants' detection ability is related across different reasoning problems. To explore this we again computed detection efficiency for every participant, but this time

[5] For our present analyses we have chosen a categorical detection index approach, i.e. participants' detection efficiency was calculated by summing up the number of times they showed the detection effect. However, some authors (e.g., Pennycook et al., 2015) favor a continuous approach to the conflict detection measurement based on the size of the detection effect. Therefore, to please all readers regardless of the approach they prefer, all of the subsequent analyses were also repeated with detection effect sizes instead of detection efficiency indices. The results are presented in the supplementary material (see Tables S4, S5, and S6). Note that despite the external dissimilarity of the two approaches, by and large, they point to very similar conclusions.

separately for every type of reasoning problem on which the participant was biased. Detection efficiency was calculated as the number of detected conflicts in a given task based on the response latency and confidence index divided by the number of detection indices. Correlation analysis showed that these indices were mostly unrelated, with detection efficiency in the bat-and-ball task showing no relation to detection efficiency in the base-rate, $r(219) = .059$, $p = .38$, or conjunction fallacy task, $r(242) = –.001$, $p = .99$. The efficiencies on the latter two were also not correlated, $r(278) = .044$, $p = .47$. The only significant correlation was between detection efficiency in syllogisms and base-rate problems, $r(225) = .133$, $p = .047$. But again, the former was unrelated both to the bat-and-ball task, $r(212) = .031$, $p = .65$, and conjunction items, $r(247) = –.053$, $p = .41$. Hence, consistent with the findings of Frey and De Neys (2017), we have found very little evidence for the domain generality of conflict detection. Thus, even if people are quite successful in detecting conflicts within one type of a reasoning problem, this ability does not necessarily seem to transfer to another type of reasoning problem. For an interested reader, we also computed correlations between successful detection observed on every reasoning problem and every detection index separately, which can be found in the supplementary materials.

Note that for completeness, we also calculated reliability estimates for the key detection efficiency indices based on the response latency and confidence index. Results indicated that these were very low (response latency index: $\alpha = .03$, confidence index: $\alpha = .14$, detection efficiency for both indices combined: $\alpha = .25$). This undoubtedly reflects the nature of conflict detection measurement and low domain generality of detection ability[6].

Predicting individual differences in detection efficiency and conflict reasoning accuracy: correlations

We now move to the critical question of how the standard individual difference predictors and mindware instantiation (i.e. average accuracy on neutral versions of reasoning problems) are related to participant's conflict detection efficiency and overall reasoning

[6] But reliability estimates are presumably also low because participants differed in the number of indices (i.e. number of tasks on which they were biased). Only those participants who were biased on all tasks could be included in the analysis ($n = 176$). We note that caution is required in interpreting indices with very low reliability, which is also likely a cause of the lower predictive power of these indices in our regression analyses. We come back to this issue in the discussion.

accuracy. Although the detection efficiencies based on response latency and confidence were correlated, the relationship was not very strong. To avoid spurious conclusions we therefore decided to run the individual difference predictor analyses separately for the two detection measures, rather than to combine them into a single index.

First, we have examined the correlations between the two detection efficiencies, standard individual difference predictors, and mindware instantiation. Table 6 gives an overview of the correlation analyses. As the table shows, participants with a higher latency detection efficiency score higher on Need for Cognition, numeracy, and cognitive ability, although the relationships with this detection index were all relatively weak. Stronger correlations were found in the case of confidence detection efficiency, which was also related to participants' cognitive reflection and mindware instantiation. While both of the latter factors as well as an NFC-like thinking disposition measure were already shown to correlate with the conflict detection ability in certain reasoning problems (Frey et al., 2018; Pennycook et al., 2014, 2015), neither cognitive ability nor numeracy were previously observed to contribute to the detection efficiency. This may have been due to the fact that their relationship with conflict detection is relatively weak and remained undetected in the less highly powered previous studies.

Next, we examined whether standard individual difference predictors, mindware instantiation, and detection efficiency are related to overall conflict reasoning accuracy. For the purpose of this as well as all of the subsequent analyses, we have computed a single conflict reasoning accuracy composite score by summing the correct answers on all sixteen conflict reasoning problems (α = .87)[7,8]. Results are included in Table 6. In line with previous research (Klaczynski, 2014; Teovanović et al., 2015; Toplak et al., 2011), the conflict reasoning accuracy composite showed moderate to strong correlations with all of the standard

[7] However, the results for every type of reasoning problem separately can be found in the supplementary material

[8] Previous research has shown that correlations between various reasoning problems tend to be relatively modest (e.g. Teovanović et al., 2015) and composite scores derived from larger set of different reasoning problems tend to show low internal consistency (Toplak et al., 2011). In contrast, correlations among reasoning problem accuracies in the present study (Table S2 in the supplementary material) range between .20 – .46 and reliability for conflict reasoning accuracy composite is high. We believe this is because we did not employ large set of different reasoning problems, but rather four different types of problems with more items per problem type. This allowed creating reliable composites for every problem type which have shown enough commonality to be summed into a single composite (for a similar approach, see Klaczynski, 2014).

individual difference predictors. There was no relationship between latency detection efficiency and conflict reasoning accuracy, but both confidence detection efficiency and mindware instantiation did correlate substantially with reasoning performance.

Table 6. Correlations between latency and confidence detection efficiency, conflict reasoning accuracy composite, standard individual difference predictors, and mindware instantiation

	1.	2.	3.	4.	5.	6.	7.	8.
1. Detection efficiency: latency	1							
2. Detection efficiency: confidence	**.24**	1						
3. Conflict reasoning accuracy	.07	**.27**	1					
4. Cognitive reflection	.07	**.22**	**.47**	1				
5. Faith in Intuition	.00	–.03	**–.26**	**–.12**	1			
6. Need for Cognition	**.14**	**.10**	**.24**	**.22**	–.09	1		
7. Numeracy	**.12**	**.19**	**.48**	**.40**	**–.11**	**.38**	1	
8. Cognitive ability	**.12**	**.19**	**.46**	**.50**	**–.19**	**.25**	.48	1
9. Mindware	.04	**.22**	**.49**	**.35**	**–.15**	**.21**	**.34**	**.35**

Note. Correlations pertaining to detection efficiencies are based on 384 observations, others on 399 observations. Correlations that appear in bold are significant at $p < .05$.

While the examination of correlations between mindware instantiation and standard individual difference predictors was not among the main aims of this study, we noticed some interesting trends in this regard which we briefly mention here. The role of mindware instantiation in both conflict detection and reasoning accuracy is theoretically acknowledged (Pennycook et al., 2015; Stanovich, 2018), yet, in empirical studies it rarely shows up to substantially contribute to reasoning performance (e.g., Frey & De Neys, 2017; Frey et al., 2018). This is presumably because the participant's performance on neutral versions of reasoning problems, which is used as a proxy for their mindware, is usually very high. In the present study, however, while the average performance on neutral problems was also high, it was strongly correlated with the conflict reasoning accuracy composite and was also related to detection efficiency based on confidence. Moreover, mindware instantiation also showed moderate correlations with all of the standard individual difference measures. Such results suggest that even though the variability in available mindware may not be very large among educated adults, individual differences in mindware instantiation may still play a non-negligible role in detection efficiency and conflict reasoning accuracy and may be a more important contributor to the reasoning performance than previously reckoned.

Predicting individual differences in detection efficiency: regression models

We now turn to our key analyses. While mindware instantiation and several standard individual difference predictors were related to the latency and confidence detection efficiency indexes, these variables were themselves all moderately intercorrelated. Thus, to determine which of these factors are the strongest independent predictors of conflict detection, we conducted two regression analyses separately for latency and confidence detection efficiency. The results are summarized in Table 7. In the first regression, NFC showed up to be the only significant predictor (β = .11) of the latency detection efficiency index when all other standard individual difference predictors and mindware instantiation were taken into account. However, the proportion of explained variance was very small (1.4%) and the overall model was only marginally significant.

Table 7. Summary of the regression analysis predicting latency and confidence detection efficiency

	Detection efficiency: latency		Detection efficiency: confidence	
	β	p	β	p
Constant		.001		.316
Mindware	–.02	.706	**.14**	**.011**
Cognitive ability	.08	.226	.05	.404
Numeracy	.05	.442	.07	.243
Need for Cognition	**.11**	**.048**	.01	.858
Faith in Intuition	.03	.558	.02	.661
Cognitive reflection	.00	.964	**.12**	**.040**
	R^2 = .01, $F(6,377)$ = 1.91, p = .078		R^2 = .07, $F(6,377)$ = 5.61, p < .001	

Note. The table contains standardized regression coefficients (β) with their respective significance. R^2 denotes adjusted r-square for the model with appropriate F-statistics. Significant regression coefficients are presented in bold.

In the second regression, both cognitive reflection (β = .12) and mindware instantiation (β = .14) predicted confidence detection efficiency after accounting for other variables in the regression. While the predictors explained more variance in case of the confidence detection index (7%), their contributions were relatively weak and leave a lot of space for other potential predictors of the conflict detection ability. Interestingly, even though Need for Cognition, numeracy, and cognitive ability were related to confidence detection efficiency in the correlation analysis above, they did not show up as significant independent predictors in the regression model. Thus, their contribution to conflict detection might be primarily caused by their relationship with cognitive reflection, which has been shown to tap all three of the abovementioned factors (e.g., Thomson & Oppenheimer, 2016).

Predicting individual differences in conflict reasoning accuracy: regression model

While the correlations presented earlier showed that detection efficiency, mindware instantiation, and standard individual difference predictors are all related to the overall accuracy in reasoning problems, we also examined these variables as independent predictors in a linear regression on the conflict reasoning accuracy composite. In the first step of the regression, we entered all standard individual difference predictors. Then, we included mindware instantiation, and at the final step we entered both detection efficiency indices to the regression model. This approach was chosen to examine whether mindware instantiation and detection efficiency, both theoretically important determinants of bias susceptibility (Pennycook et al., 2015; Stanovich, 2018), predict reasoning accuracy over and above cognitive ability, thinking dispositions, numeracy, and cognitive reflection measures. The results are summarized in Table 8.

All variables in the final model except for Need for Cognition and latency detection efficiency did significantly contribute to conflict reasoning accuracy. Among the standard individual difference predictors, numeracy (β = .22), cognitive reflection (β = .19), Faith in Intuition (β = –.14), and cognitive ability (β = .11) were found to independently predict the accuracy on conflict problems at the last step of the model. Overall, these standard individual difference predictors explained 34% of the variance in the conflict reasoning accuracy composite. Our results in this respect are consistent with previous individual difference predictor analyses (Klaczynski, 2014; Toplak et al., 2011). More critically, we examined whether mindware instantiation and conflict detection efficiency play a role in reasoning accuracy over and above standard individual difference predictors. At the second step of the regression, mindware instantiation accounted for 5% of additional variance over the standard individual difference measures and ended up being the strongest independent predictor (β = .23) of conflict reasoning accuracy composite in the final model. Lastly, while the predictive power of confidence detection efficiency was not overly strong (β = .11), it did show up as an independent predictor at the final step of the regression and it accounted for another 1% of the variance in accuracy on conflict problems after all of the other variables were accounted for. Thus, despite the substantial proportion of variance in conflict reasoning accuracy which was already explained by standard individual difference predictors, both mindware instantiation and confidence detection efficiency further contributed to the reasoning performance, in line with their hypothesized distinct theoretical role as key determinants of bias susceptibility.

Together the variables explained approximately 40% of the variance in the conflict reasoning accuracy composite.

Table 8. Summary of the regression analysis predicting the composite of correctly answered conflict reasoning problems

	β	p
Step 1		
Constant		< .001
Cognitive ability	**.15**	**.004**
Numeracy	**.27**	**.001**
Need for Cognition	.03	.524
Faith in Intuition	**–.15**	**< .001**
Cognitive reflection	**.26**	**< .001**
R^2 = .34, F(5,378) = 40.98, p < .001		
Step 2		
Constant		.690
Cognitive ability	**.11**	**.026**
Numeracy	**.23**	**< .001**
Need for Cognition	.01	.763
Faith in Intuition	**–.13**	**.001**
Cognitive reflection	**.21**	**< .001**
Mindware	**.25**	**< .001**
ΔR^2 = .05, F(1,377) = 31.67, p < .001		
Step 3		
Constant		.733
Cognitive ability	**.11**	**.031**
Numeracy	**.22**	**< .001**
Need for Cognition	.01	.745
Faith in Intuition	**–.14**	**.001**
Cognitive reflection	**.19**	**< .001**
Mindware	**.23**	**< .001**
Detection efficiency: LAT	–.02	.640
Detection efficiency: CON	**.11**	**.008**
ΔR^2 = .01, F(2,375) = 3.59, p = .029		

Note. The table contains standardized regression coefficients (β) with their respective significance. R^2 and ΔR^2 denote adjusted r-square for the initial model and change in r-square at the 2nd and 3rd step of the regression with appropriate change statistics. LAT: latency, CON: confidence. Significant regression coefficients are presented in bold.

DISCUSSION

In the present study, we set out to examine individual differences in the ability to detect intuition/logic conflict and people's overall reasoning accuracy while employing several traditional reasoning problems and detection indices to ensure the robustness of our results. We found that the Need for Cognition thinking disposition, mindware instantiation, and cognitive reflection are independent predictors of participants' detection ability, although the overall explained variance was quite low. Our results also show that detection efficiency and mindware instantiation are both predictors of the accuracy on conflict reasoning problems over and above the measures of cognitive ability, thinking dispositions, numeracy, and cognitive reflection. This is consistent with the hypothesized theoretical role of conflict detection and mindware instantiation as essential processes in the susceptibility to cognitive biases (De Neys & Bonnefon, 2013; Pennycook et al., 2015; Stanovich, 2018).

Our key finding is that although various standard individual difference predictors played a significant role, the single best predictor of both conflict detection efficiency and overall reasoning accuracy turned out to be mindware instantiation. Theoretically, the availability of the necessary mindware that allows one to grasp the normative solution of the task at hand is thought to be one of the key factors for successful conflict detection as well as overall bias susceptibility (De Neys & Bonnefon, 2013; Stanovich, 2018). Yet, mindware instantiation is often neglected in empirical studies on rational thinking because it is presumed that most educated adults do possess the basic rules of logic and mathematics necessary for solving the traditional reasoning tasks. This assumption seems to be supported by the almost perfect accuracy on neutral versions of such problems (e.g., De Neys & Glumicic, 2008; Frey & De Neys, 2017; Frey et al., 2018). While we have also observed high average performance (around 80%) on neutral reasoning tasks, our results showed that despite the low variability, individual differences in mindware instantiation were still strongly related to the overall conflict reasoning accuracy, and to a more moderate extent with one's conflict detection ability.

The link between conflict detection, mindware instantiation, and conflict reasoning accuracy may not be that surprising as the three factors are all indexed by very similar tasks, i.e., conflict, no-conflict, and neutral versions of the same reasoning problems. However, mindware instantiation was also moderately related to all of the standard individual difference predictors. This further supports the view that the differences between participants in

available mindware are indeed meaningful, even if the variations in neutral reasoning problem accuracy are not large. Taken together, our results suggest that mindware instantiation may be a more important source of individual differences in conflict detection and overall reasoning accuracy than previously thought, and it should receive more attention in research on cognitive biases (see also Stanovich, 2018).

Along with mindware instantiation, several standard individual difference predictors contributed consistently to both detection efficiency and overall conflict reasoning accuracy. Their relationship with conflict detection ability in the present study was by and large consistent with the partial results presented by Pennycook et al. (2014, 2015), who have shown that both cognitive reflection and thinking dispositions are related to detection ability in the base-rate neglect task. Both of the aforementioned variables can be thought of as indicators of the propensity to recognize when one's intuitive thinking may be insufficient and more effortful processing is needed (Frederick, 2005; Stanovich et al., 2016), which might explain their contribution to one's conflict detection ability. Also, in line with Swan et al. (2018; however see Thompson & Johnson, 2014), we have observed a weak correlation between cognitive ability and the confidence and latency detection efficiencies, but this relationship did not hold in the regression where other predictors were taken into account.

Whereas standard individual difference predictors played a relatively modest role in the conflict detection ability, they had a much more significant contribution to participant's overall conflict reasoning accuracy. Our results in this regard again concur with other studies which simultaneously examined several variables related to reasoning performance and found that cognitive ability, thinking dispositions, and numeracy or cognitive reflection are all independent predictors of conflict reasoning accuracy (Klaczynski, 2014; Toplak et al., 2011). The present research, however, brings a more comprehensive analysis which also takes into account estimates of participants' conflict detection ability and mindware instantiation, which showed up to predict reasoning over and above the standard individual difference predictors. Together, the full regression model accounted for 40% of the variance in conflict reasoning accuracy. This result, however, stands in sharp contrast with the predictive power of the regression models pertaining to conflict detection which only accounted for 7% of the variance in confidence detection efficiency, and even less in the efficiency index based on participant's response latencies. We discuss possible reasons for this difference later below.

When analyzing individual differences in conflict detection, we have found some inconsistencies in the results pertaining to different detection efficiency indices. By and large, the index based on response latencies consistently yielded far less clear patterns of results than the one based on confidence. In comparison with the confidence index, the latency detection efficiency was not significantly related to mindware instantiation (i.e., the most potent predictor in the present research), its correlations with standard individual difference predictors were generally low, and the regression model with all variables explained only 2% of variance in this detection efficiency index. A possible explanation for this is the relative noisiness of the response latency measure. While reasoning problems were presented to participants in several parts to disentangle reading and decision latency as much as possible (e.g., Frey et al., 2018), the timing on the tasks was not restricted. This, together with the fact that wording of some reasoning problems is still quite lengthy (e.g., Pennycook et al., 2015), could result in larger variations in response time measurement rendering this method noisy. Also, as can be seen from the analyses presented in the supplementary materials, the results pertaining to the confidence latency measure were quite distinct from the other two detection indices. It was not related to any of the standard individual difference predictors, mindware instantiation, nor accuracy on conflict reasoning problems. Moreover, it showed almost no relation to the two other detection efficiencies, and also generally produced weaker detection effects than response latency and confidence measures. Taken together with the conclusions presented by Frey et al. (2018), these results strongly suggest that confidence latency is not a reliable indicator of the detection ability. Therefore, we would like to warn researchers to be careful when employing this measure in future conflict detection studies.

In line with previous studies, the present results also point to the limited domain generality of conflict detection (Frey & De Neys, 2017; Frey et al., 2018). We observed quite some variability in the conflict detection efficiency across various reasoning problems. This means that even if someone is capable of registering conflicts on certain problems, they cannot be expected to also show more successful detection in other tasks. The low generality of conflict detection ability may well be the key reason for why we also obtained very low internal consistency estimates of conflict detection indices. Despite our effort to use the most robust way of measuring individual differences in conflict detection ability by employing four different reasoning problems and three detection indices, resulting detection efficiency estimates were far below satisfactory reliability. This certainly means that caution is needed

when interpreting the results of conflict detection analyses and that our results will need to be replicated before drawing strong conclusions.

And yet, observed relationships between conflict detection indices, reasoning performance, and standard individual predictors, were quite in line with both previous partial research findings (Frey et al., 2018; Pennycook et al., 2014, 2015), and theoretical predictions regarding detection ability as important component of bias susceptibility (De Neys & Bonnefon, 2013; Pennycook et al., 2015; Stanovich, 2018). Also, the fact that detection efficiencies did show up to predict conflict reasoning accuracy over standard individual difference predictors and mindware instantiation suggests that detection indices did capture some meaningful variance which is predictive of reasoning performance and is not due to other related cognitive factors which were included in the analysis. Most importantly, a recent study by Burič and Šrol (2019) offers some evidence for the replicability of the present conflict detection findings and thus lends further credence to the results we report here. Although their research was not designed as a direct replication of our study, the authors measured conflict detection as we did here. Their results show similar patterns of individual differences in conflict detection efficiency related to cognitive reflection and mindware instantiation and replicate the predictive role of detection efficiency and mindware instantiation in reasoning accuracy observed on conflict syllogisms and base-rate neglect tasks.

It should be noted, however, that Burič and Šrol (2019) also identified the problem with the low reliability of the conflict detection indices. We believe this low reliability is the key reason for why we have only managed to explain 7% of the variance in the confidence detection efficiency, even though we used a wide range of standard individual difference predictors as well as mindware instantiation as possible predictors. Given the low intercorrelations of detection efficiency between specific reasoning domains, any model examining general detection ability across tasks probably should not be expected to explain too much of the common variance. In the same way, low internal consistency has to be borne in mind when considering why some of the conceptually relevant variables did not show any relationship or were only modestly correlated with conflict detection, especially in the analyses pertaining to the latency detection index.

Taking into the account the low reliability of the conflict detection indices, relatively low correlations between detection effects observed across different tasks and measures, and

overall low observed effects in the conflict detection analyses, it seems that further research on individual differences in the conflict detection mechanism will face a challenging task to overcome the issues identified in the present study. Still, it is clear that individual differences in conflict detection should not be disregarded, despite the problems the measurement of these differences may bear.

At this point, we would like to discuss some more theoretical implications of our findings for ongoing developments in the field of dual process theories of reasoning. As we mentioned, the conflict detection findings have led to a reformulation of traditional reasoning theories by positing that type 1 processing cues multiple intuitive responses based both on heuristics and the logical structure of the task (De Neys, 2012, 2013, 2017; Pennycook et al., 2015; Thompson & Newman, 2017). Thus, what people are actually detecting in reasoning tasks is a conflict between two competing types of intuitions – one heuristic and the other logical. As such, the conflict detection mechanism is viewed under the recent hybrid dual process models to be a result of type 1 processing (De Neys, 2017; Pennycook, et al., 2015). Further, the likelihood of conflict detection and subsequent engagement in analytic type 2 processing has been thought to be determined by the relative strength of the logical and heuristic intuition. Specifically, it is assumed that conflict detection likelihood will be maximal when the strength of the two intuitively cued outputs is maximally similar (Bago & De Neys, 2019a; De Neys & Pennycook, 2019; Pennycook et al., 2015). However, as for most biased reasoners the heuristic intuition will be typically stronger than the logical one, correct responding on conflict reasoning tasks for them will require the analytic type 2 processing to override the dominant heuristic intuition.

One striking discrepancy in our results is that our predictors explained a much larger proportion of the variance in reasoning accuracy than in conflict detection. This may be due to the fact that reasoning accuracies observed on different problems are at least moderately intercorrelated (see table S2 in the supplementary material), which suggest more domain generality for overall reasoning accuracy than for conflict detection ability per se. In theoretical terms, this might imply that the type 2 process override success or capacity is more invariant across reasoning tasks than the type 1 detection component of the model. As reasoning performance is presumed to depend not only on efficient detection, but also intuition inhibition (De Neys & Bonnefon, 2013; Pennycook et al., 2015; Stanovich, 2018), it might be that the latter process is more domain general than the former.

More specifically, given that the conflict detection likelihood is assumed to depend on the relative strength of one's heuristic and/or logical intuitions (Pennycook et al., 2015; Bago & De Neys, 2019a; De Neys & Pennycook, 2019), these relative strengths might show much more variability across different reasoning tasks and/or even across different item contents of a single task. As is clear from the traditional group and individual-level conflict detection analyses both in the present study and in previous works (e.g., Frey & De Neys, 2017; Frey et al., 2018), there is quite some variability in detection effects observed across different tasks. Moreover, even relatively minor changes to the specific features within the same reasoning task, such as manipulating the extremity of base-rates or the overall order of presented information within the reasoning problem, have been shown to influence the likelihood of conflict detection (Pennycook et al., 2012; 2015), presumably by differently increasing the strength of either the heuristic or logical intuition. Obviously, if the likelihood of conflict detection is determined by the difference in strength between the logical and heuristic intuitions and the strength of those intuitions varies even with small changes within the reasoning task, it is unlikely that conflict detection effectiveness would be very stable across problems with different formats and item contents. We speculate that it might be precisely this sensitivity of type 1 processing which leads to low domain generality of conflict detection indices and the associated methodological problems identified in the present research.

Although speculative, we believe that the difference between more context dependent intuitive type 1 processing in comparison with more domain general type 2 processing may help to account for the large observed gap between explained variance in detection efficiency and overall reasoning accuracy. Of course, this also suggests that currently it will be hard to study individual predictors of conflict detection, as researchers will have to deal with a substantial within and across task variability stemming from the type 1 processing. While we cannot currently offer any definitive solutions for the problems presented here, we hope that the analysis might still help to raise awareness of the issues that might complicate future research on individual differences in conflict detection.

Certainly, our study is not without its limitations. For one, our mindware instantiation index, which turned up to be the most consistent predictor in the present study, was of poor reliability. As with conflict detection, mindware instantiation has been argued to be considerably subject and task specific (Stanovich, 2018). This, coupled with the participant's ceiled performance on neutral reasoning problems may have led to the poor reliability of the mindware instantiation measure. Nevertheless, we have observed stable patterns of

correlations between mindware instantiation and both conflict reasoning accuracy and standard individual difference predictors. As a side note, this may be also due to the fact that Cronbach's α represents an estimate of the lower bound of reliability of a given measure (Mair, 2018). Still, the low reliability of the mindware instantiation index points to the need to replicate the present results. It would be worthwhile to dedicate further effort to try to come up with a more reliable approach to measure participant's mindware instantiation, which would allow to better study this neglected component of cognitive bias research (Stanovich, 2018).

As one reviewer noted, since one of the main problems with our mindware instantiation measure lied in participants' very high average performance, one possible solution would be to increase the difficulty of neutral reasoning problems. However, we believe that such a solution comes with its own limitations. Specifically, while more complex neutral problems may better tap participants' potential, they will no longer represent the specific mindware needed to solve the less complex no-conflict and conflict versions. Still, it is possible that a simple solution such as increasing the number of neutral items might help researchers in future studies to limit the problem of low internal consistency and thus increase the reliability of findings pertaining to the role of the mindware component in the reasoning process. Although the mindware instantiation in our study showed up to be the most substantial predictor of both conflict detection and overall accuracy on conflict reasoning problems, due to low reliability of our mindware measure, the results will have to be taken with some caution (although, for similar patterns of results pertaining to mindware instantiation, see Burič & Šrol, 2019).

Secondly, in our choice of standard individual predictor measures, we have relied on relatively short tests intended to tap the constructs of interest with only several items. This was done mainly because we wanted to reduce participants' fatigue resulting from the length of study (which was already exacerbated by the fact that we employed four types of problems to study reasoning accuracy and conflict detection). Future research might want to supplement our choice of predictor measures with longer, more reliable scales, as well as methods tapping into other constructs which have been previously shown to predict conflict reasoning accuracy, such as verbal intelligence, and/or actively open-minded thinking disposition (Stanovich et al., 2016).

Individual difference studies are an integral part of the research on cognitive biases and have been paramount in advancing our understanding of the processes which are implicated in sound reasoning and decision-making (Stanovich & West, 2008; Stanovich et al., 2016; Teovanović et al., 2015). While up till now the studies have uncovered several standard individual difference predictors which independently predict accuracy on conflict reasoning problems (e.g., Klaczynski, 2014; Toplak et al., 2011), they failed to relate these factors to specific components of bias susceptibility (De Neys & Bonnefon, 2013; Pennycook et al., 2015; Stanovich & West, 2008). In the present study, we set out to fill this gap by examining individual difference predictors specifically related to one of these components, the ability to detect a conflict between intuition and logic, and determine how these differences relate to overall accuracy on reasoning problems. We have found that while several standard individual difference predictors contributed to various extent to both conflict detection and reasoning accuracy, the most important factor in both regards showed up to be mindware instantiation. Mindware instantiation has long been recognized as a theoretically critical component of sound reasoning (Stanovich, 2018; Stanovich & West, 2008). However, up until now, it was mostly neglected in the empirical research in this area. Thus, the present study highlights the importance of teasing apart specific components of sound reasoning and studying their relative contributions to overall susceptibility to cognitive biases.

DECLARATION OF INTEREST

None.

ACKNOWLEDGEMENTS

The study is based on an unpublished doctoral dissertation of the first author (J.Š.). We would like to thank Maggie Toplak, Keith Stanovich, Valerie Thompson, and two anonymous reviewers for their helpful comments on an earlier draft of this manuscript. This study was supported by the Slovak Research and Development Agency and is part of the research project APVV-16-0153: "*Cognitive failures – individual predictors and intervention possibilities*". We would also like to thank the ANR for their support (ANR-16-CE28-0010-01). Data for this study are publicly available at OSF: https://osf.io/3cp9u/

REFERENCES

Bago, B., & De Neys, W. (2019a). Advancing the specification of dual process models of higher cognition: A critical test of the hybrid model view. *Thinking & Reasoning*, http://dx.doi.org/10.1080/13546783.2018.1552194.

Bago, B., Frey, D., Vidal, J., Houdé, O., Borst, G., & De Neys, W. (2018). Fast and slow thinking: Electrophysiological evidence for early conflict sensitivity. *Neuropsychologia*, *117*(June), 483–490.

Burič, R., & Šrol, J. (2019). Individual differences in logical intuitions on reasoning problems presented under two-response paradigm. https://doi.org/10.31234/osf.io/w6un2

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*(1), 25–47.

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science*, *7*(1), 28–38.

De Neys, W. (2013). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning*, *20*, 1–19.

De Neys, W. (2017). Bias, conflict, and fast logic: towards a hybrid dual process future? In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 47–65). London: Routledge.

De Neys, W., & Bonnefon, J.-F. (2013). The "whys" and "whens" of individual differences in thinking biases. *Trends in Cognitive Sciences*, *17*(4), 172–178.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248–1299.

De Neys, W., Moyens, E., & Vansteenwegen, D. (2010). Feeling we're biased: autonomic arousal and reasoning conflict. *Cognitive, Affective & Behavioral Neuroscience*, *10*(2), 208–216.

De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, 0963721419855658.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*(2), 269–73.

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter Than We Think. *Psychological*

Science, *19*(5), 483–489.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390–405.

Evans, J. S. B. T. (2007). On the resolution of conflict in dual process theories of reasoning. *Thinking & Reasoning*, *13*(4), 321–339.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science*, *8*(3), 223–241.

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. *Medical Decision Making*, *27*(5), 672–680.

Franssens, S., & De Neys, W. (2009). The effortless nature of conflict detection during thinking. *Thinking & Reasoning*, *15*(2), 105–128.

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Frey, D., & De Neys, W. (2017). Is Conflict Detection in Reasoning Domain General ? In *Proceedings of the Annual Meeting of the Cognitive Science Society, 39* (pp. 391–396).

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *The Quarterly Journal of Experimental Psychology*, *71*(5), 1188–1208.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Penguin Books.

Klaczynski, P. A. (2014). Heuristics and biases: Interactions among numeracy, ability, and reflectiveness predict normative responding. *Frontiers in Psychology*, *5*(JUL), 1–13.

Klose, J., Černochová, D., & Král, P. (2002). *Vídeňský maticový test*. Praha: Testcentrum.

Mair, P. (2018). *Modern Psychometrics with R*. New York: Springer.

Mata, A., Ferreira, M. B., Voss, A., & Kollei, T. (2017). Seeing the conflict : an attentional account of reasoning errors. *Psychonomic Bulletin & Review*, *24*(6), 1980–1986.

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houdé, O., & De Neys, W. (2014).

Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology*, *27*(2), 227–237.

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*(1), 101–6.

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.

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. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*. http://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–695.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The Rationality Quotient: Toward a test of rational thinking*. MIT Press.

Stupple, E. J., Ball, L. J., & Ellis, D. (2013). Matching bias in syllogistic reasoning: Evidence for a dual- process account from response times and confidence ratings. *Thinking & Reasoning*, *19*(1), 54–77.

Stupple, E. J., & Ball, L. J. (2008). Belief-Logic Conflict Resolution in Syllogistic Reasoning: Inspection-Time Evidence for a Parallel-Process Model. *Thinking Reasoning*, *14*(2), 168–181.

Swan, A. B., Calvillo, D. P., & Revlin, R. (2018). To detect or not to detect: A replication and extension of the three-stage model. *Acta Psychologica*, *187*(October 2017), 54–65.

Šrol, J. (2019). *Individual differences in susceptibility to cognitive biases: implication for theories of rational thought*. Unpublished doctoral thesis, Slovak Academy of Sciences, Bratislava, SK.

Teovanović, P., Knežević, G., & Stankov, L. (2015). Individual differences in cognitive biases: Evidence against one-factor theory of rationality. *Intelligence*, *50*, 75–86.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking & Reasoning*, *20*(January 2015), 215–244.

Thompson, V. A., & Newman, I. R. (2017). Logical intuitions and other conundra for dual process theories. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 121–136). London: Routledge.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*(7), 1275–1289.

Tversky, A., & Kahneman, D. (1983). Extensional Versus Intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Vartanian, O., Beatty, E. L., Smith, I., Blackler, K., Lam, Q., Forbes, S., & De Neys, W. (2018). The Reflective Mind: Examining Individual Differences in Susceptibility to Base Rate Neglect with fMRI. *Journal of Cognitive Neuroscience*, *30*(7), 1–12.

SUPPLEMENTARY MATERIAL

Section A

Individual-level conflict detection analyses separately for every type of reasoning problem

Following Frey et al. (2018), separately for every detection measure and type of reasoning problem, participants who got at least one conflict item incorrect and one no-conflict item correct (whole biased group) were further divided into three subgroups, according to whether they showed longer latencies[9] and lower confidence for incorrect conflict than correct no-conflict problems (*detection subgroup*), the opposite pattern of results (*reverse detection*), or the same latency and confidence estimates for the two versions of problems (*same subgroup*). The proportions of participants in the three detection subgroups and the whole biased group, as well as the detection effects for every detection index (i.e. the difference in response latency, confidence, or confidence latency between incorrectly solved conflict and correctly solved no-conflict problems) and correlations between the size of the detection effect and accuracy on conflict problems separately for the three subgroups and the whole biased group are summarized in the Table S1. Note that for correlation analyses the detection effects were reversed so that higher numbers represent better detection (i.e. larger response latency increase or confidence decrease to the conflict in comparison with no-conflict tasks).

Across various indices, 29 – 79% of biased reasoners showed signs of successful conflict detection. The relationships between detection effects and accuracies were all positive in the detection subgroup, however, they have ranged from very low, mostly in the case of confidence latency index in conjunction fallacy problems (r = .01, *n.s.*), to quite strong, the latter being true mostly in the case of confidence detection effect in the syllogisms (r = .42), bat-and-ball items (r = .39), and base-rate neglect tasks (r = .51). Among the whole sample of biased reasoners correlations between the size of the detection effect and the accuracy on conflict problems were generally weaker than in the detection subgroup and not all of them showed up to be statistically significant.

[9] Prior to the analyses, all latency data were checked for outlying observations. Latency values which were more than three standard deviations above/below the mean of the respective index were replaced with the value of three standard deviations above/below the average. All analyses reported in the supplementary material are based on the outlier treated data. However, all analyses were also run on the raw data (before outlier replacement) and the results were consistent with the conclusions presented in the study.

Table S1. The results of the conflict detection analyses based on the three detection measures for three subgroups and the whole biased reasoners group separately for every type of reasoning problem

	Subgroup detection	Reverse detection	Subgroup same	Whole biased group
Syllogistic reasoning task				
Response latency (% of biased group)	153 (57%)	117 (43%)	–	270 (100%)
conflict detection effect (*SD*)	–2.75 (3.38)	1.89 (2.07)	–	–0.74 (3.69)
detection – accuracy *r (p)*	**.22 (.006)**	.04 (.690)	–	**.14 (.021)**
Confidence (% of biased group)	97 (36%)	94 (35%)	79 (29%)	270 (100%)
conflict detection effect (*SD*)	–1.50 (1.33)	1.16 (1.14)	0	–0.13 (1.53)
detection – accuracy *r (p)*	**.42 (.001)**	–.05 (.615)	–	**.24 (.001)**
Confidence latency (% of biased group)	146 (54%)	124 (46%)	–	270 (100%)
conflict detection effect (*SD*)	–0.99 (1.44)	0.73 (0.92)	–	–0.20 (1.50)
detection – accuracy *r (p)*	**.29 (.001)**	.03 (.739)	–	**.15 (.012)**
Bat-and-ball items				
Response latency (% of biased group)	208 (76%)	65 (24%)	–	273 (100%)
conflict detection effect (*SD*)	–5.78 (7.11)	2.85 (3.54)	–	–3.72 (7.42)
detection – accuracy *r (p)*	**.14 (.044)**	–.07 (.590)	–	.10 (.109)
Confidence (% of biased group)	79 (29%)	18 (7%)	176 (64%)	273 (100%)
conflict detection effect (*SD*)	–2.16 (2.43)	0.75 (0.58)	0	–0.58 (1.66)
detection – accuracy *r (p)*	**.39 (.001)**	.10 (.702)	–	**.30 (.001)**
Confidence latency (% of biased group)	152 (56%)	121 (44%)	–	273 (100%)
conflict detection effect (*SD*)	–1.13 (1.77)	0.76 (1.38)	–	–0.29 (1.86)
detection – accuracy *r (p)*	**.32 (.001)**	.05 (.557)	–	**.19 (.002)**
Base-rate neglect				
Response latency (% of biased group)	169 (57%)	125 (43%)	–	294 (100%)
conflict detection effect (*SD*)	–4.97 (6.25)	2.41 (2.70)	–	–1.83 (6.23)
detection – accuracy *r (p)*	**.23 (.003)**	–.05 (.596)	–	**.23 (.001)**
Confidence (% of biased group)	188 (64%)	88 (30%)	18 (6%)	294 (100%)
conflict detection effect (*SD*)	–2.05 (1.60)	0.88 (0.62)	0	–1.04 (1.89)
detection – accuracy *r (p)*	**.51 (.001)**	**–.24 (.025)**	–	**.39 (.001)**
Confidence latency (% of biased group)	140 (48%)	154 (52%)	–	294 (100%)
conflict detection effect (*SD*)	–0.72 (0.85)	0.70 (0.78)	–	0.02 (1.08)
detection – accuracy *r (p)*	.13 (.117)	.03 (.762)	–	.10 (.104)
Conjunction fallacy problems				
Response latency (% of biased group)	268 (79%)	73 (21%)	–	341 (100%)
conflict detection effect (*SD*)	–4.44 (4.37)	2.72 (3.23)	–	–2.91 (5.08)
detection – accuracy *r (p)*	**.24 (.001)**	–.19 (.115)	–	**.13 (.018)**
Confidence (% of biased group)	257 (75%)	55 (16%)	29 (9%)	341 (100%)
conflict detection effect (*SD*)	–1.67 (1.29)	0.80 (0.61)	0	–1.13 (1.50)
detection – accuracy *r (p)*	**.15 (.017)**	**–.47 (.001)**	–	.06 (.262)
Confidence latency (% of biased group)	168 (49%)	173 (51%)	–	341 (100%)
conflict detection effect (*SD*)	–0.59 (0.60)	0.76 (0.82)	–	0.10 (0.99)
detection – accuracy *r (p)*	.01 (.924)	–.08 (.281)	–	–.07 (.217)

Note. Response latency data are reported in seconds. For correlational analysis, detection effects are reversed so that positive values indicate better conflict detection ability. Significant detection – accuracy correlations are presented in bold.

Section B

Correlations between detection efficiencies, accuracies on individual conflict reasoning problems, standard individual difference predictors, and mindware instantiation

The correlation analysis summarized in Table 5 in the main manuscript presents correlations between standard individual difference predictors, mindware instantiation, conflict detection efficiencies, and conflict reasoning accuracy composite. For completeness, we include here the same correlation analysis with accuracies for individual conflict problems presented separately. As can be seen from Table S2, the correlations for individual reasoning problems in most regards reflect the results observed with the overall conflict reasoning accuracy composite. The most notable exceptions are that both base-rate neglect accuracy and conjunction fallacy accuracy are not correlated with confidence detection efficiency and NFC.

Table S2. Correlations between response latency and confidence detection efficiency, accuracy on conflict problems, standard individual difference predictors, and mindware instantiation

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Detection efficiency: latency	1											
2. Detection efficiency: CON	**.24**	1										
3. Syllogisms ACC	.07	**.29**	1									
4. Bat-and-ball ACC	.01	**.35**	**.46**	1								
5. Base-rate neglect ACC	.09	.08	**.38**	**.30**	1							
6. Conjunction fallacy ACC	.03	–.03	**.29**	**.20**	**.43**	1						
7. Conflict reasoning ACC	.07	**.27**	**.75**	**.72**	**.73**	**.64**	1					
8. Cognitive reflection	.07	**.22**	**.38**	**.52**	**.28**	**.13**	**.47**	1				
9. Faith in Intuition	.00	–.03	**–.23**	**–.14**	**–.19**	**–.17**	**–.26**	**–.12**	1			
10. Need for Cognition	**.14**	**.10**	**.28**	**.25**	.05	.07	**.24**	**.22**	–.09	1		
11. Numeracy	**.12**	**.19**	**.39**	**.44**	**.28**	**.22**	**.48**	**.40**	**–.11**	**.38**	1	
12. Cognitive ability	**.12**	**.19**	**.42**	**.48**	**.22**	**.16**	**.46**	**.50**	**–.19**	**.25**	.48	1
13. Mindware	.04	**.22**	**.36**	**.36**	**.37**	**.27**	**.49**	**.35**	**–.15**	**.21**	**.34**	**.35**

Note. Correlations pertaining to detection efficiencies are based on 384 observations, others on 399 observations. Correlations that appear in bold are significant at $p < .05$. CON: confidence, ACC: accuracy.

Section C

Analyses pertaining to the confidence latency detection efficiency

As the patterns of results pertaining to the confidence latency index, specifically low overall detection effect, weak and mostly insignificant correlations between the size of the detection and accuracy on conflict reasoning problems and weak correlation between confidence latency and the other two detection indices suggested it does not serve as a particularly sensitive conflict detection measure, we have decided to drop it from individual difference analyses included in the main manuscript. However, we present the results pertaining to the confidence latency detection efficiency here for completeness.

Correlations between confidence latency detection efficiency, accuracy on conflict problems, standard individual difference predictors, and mindware instantiation

In line with the results presented in the main paper, reliability estimate for confidence latency detection efficiency was also very low (α = .06). Confidence latency detection efficiency was not related to any of the conflict reasoning problems or standard individual difference predictors. Its correlations with reasoning problems were all negative, for *syllogisms*, $r(384) = -.04$, $p = .48$, *bat-and-ball items*, $r(384) = -.02$, $p = .67$, *base-rate neglect tasks*, $r(384) = -.01$, $p = .78$, *conjunction fallacy problems*, $r(384) = -.04$, $p = .39$, and *conflict reasoning accuracy composite*, $r(384) = -.04$, $p = .42$, however, none of them reached significance. Considering the standard individual difference predictors, the correlations showed no consistent patterns of relationships of confidence latency detection efficiency with *Vienna matrix test*, $r(384) = -.03$, $p = .60$, *numeracy*, $r(384) = .02$, $p = .71$, *Need for Cognition*, $r(384) = -.01$, $p = .89$, *Faith in Intuition*, $r(384) = .02$, $p = .71$, and *cognitive reflection*, $r(384) = .05$, $p = .29$, but again, none of correlations were statistically significant. Confidence latency detection efficiency was also uncorrelated with mindware instantiation, $r(384) = -.00$, $p = .97$.

Individual predictors of the confidence latency detection efficiency

Linear regression summarized in Table S3 showed that neither standard individual difference predictors nor mindware instantiation significantly predicted the confidence latency detection efficiency. Moreover, the overall regression model was not significant.

Table S3. Summary of the regression analysis predicting confidence latency detection efficiency

	β	p
Constant		.001
Mindware	–.01	.837
Cognitive ability	–.07	.257
Numeracy	.03	.622
Need for Cognition	–.02	.779
Faith in Intuition	.02	.752
Cognitive reflection	.09	.162
R^2 = –.01, F(6,377) = 0.48, p = .821		

Note. The table contains standardized regression coefficients (β) with their respective significance. R^2 denotes adjusted r-square for the model with appropriate F-statistics. Significant regression coefficients are presented in bold.

Standard individual difference predictors, mindware instantiation, and confidence latency detection efficiency as predictors of conflict reasoning accuracy

Finally, to ascertain whether confidence latency detection efficiency predicted conflict reasoning accuracy composite independently from other predictors, we have included this detection index instead of the two detection indices based on response latencies and confidence at the third step of the regression which is presented in Table 7 in the main manuscript. Entering confidence latency detection efficiency did not lead to a significant increase in model fit, ΔR^2 = .003, F(1,376) = 1.63, p = .202. Moreover, standardized regression coefficient for the confidence latency monitoring efficiency was low (b = –.05, p = .202). Curiously, however, confidence latency was negatively related to the overall conflict reasoning accuracy, although the relationship was not statistically significant.

Section D

Individual difference in conflict detection analyses using detection effect sizes

In the main manuscript, we have chosen a categorical approach to analyzing individual differences in the detection ability (i.e. analyses with detection efficiencies), based on the number of times participants showed successful detection on a given index divided by the total number of reasoning tasks on which they were biased. However, we have also repeated the key analyses using a continuous approach based on detection effect sizes (see Pennycook et al., 2015). To do this, we have calculated the mean difference between incorrectly answered conflict and correctly answered no-conflict problems in response latency, confidence, and confidence latency (averaged across all reasoning problems). Below we present analyses in which we used the three detection effect sizes averaged across all reasoning problems as indicators of participants' detection ability instead of detection efficiency indices. Detection effect sizes are reversed for simplicity so that positive values indicate better conflict detection ability (i.e. larger latency increase or confidence decrease to the conflict in comparison with no-conflict tasks). As was the case with the detection efficiency indices in the main manuscript, detection effects calculated on bases of response latency (α = .25), confidence (α = .05), and confidence latency (α = .13) also showed very low reliability estimates.

As can be seen from the Table S4, confidence detection effect is positively related with all of the conflict reasoning problems, mindware instantiation, and standard individual difference predictors, with the exception of negative correlation with FI. On the other hand, both response latency and confidence latency detection effects are mostly unrelated to other variables, the exception being weak correlations between the response latency detection, NFC, cognitive ability, and accuracy on base-rate neglect items. The only significant relationship in case of confidence latency detection effect was with cognitive reflection.

Table S4. Correlations between detection effect sizes, accuracy on conflict problems, standard individual difference predictors, and mindware instantiation

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.
1. Latency detection effect	1												
2. CON detection effect	**.16**	1											
3. CON latency detection effect	.06	**.36**	1										
4. Belief bias ACC	.05	**.31**	.07	1									
5. Bat-and-ball ACC	–.01	**.25**	.07	**.46**	1								
6. Base-rate neglect ACC	**.12**	**.17**	.04	**.38**	**.30**	1							
7. Conjunction fallacy ACC	.08	**.15**	.09	**.29**	**.20**	**.43**	1						
8. Bias susceptibility composite	.09	**.33**	.10	**.75**	**.72**	**.73**	**.64**	1					
9. Cognitive reflection	.03	**.21**	**.13**	**.38**	**.52**	**.28**	**.13**	**.47**	1				
10. Intuitive thinking disp.	–.01	**–.15**	–.06	**–.23**	**–.14**	**–.19**	**–.17**	**–.26**	**–.12**	1			
11. Analytic thinking disp.	**.15**	**.13**	.04	**.28**	**.25**	.05	.07	**.24**	**.22**	–.09	1		
12. Numeracy	.09	**.24**	.09	**.39**	**.44**	**.28**	**.22**	**.48**	**.40**	**–.11**	**.38**	1	
13. Cognitive ability	**.12**	**.22**	.03	**.42**	**.48**	**.22**	**.16**	**.46**	**.50**	**–.19**	**.25**	**.48**	1
14. Mindware instantiation	.04	**.25**	.04	**.36**	**.36**	**.37**	**.27**	**.49**	**.35**	**–.15**	**.21**	**.34**	**.35**

Note. Correlations involving detection effects are based on 384 observations, others are based on 399 observations. Correlations that appear in bold are significant at $p < .05$. CON: confidence, ACC: accuracy.

Table S5 presents independent predictors of the three detection effects. Among the three regression models, only the one predicting confidence detection effect emerged as significant with 9% explained variance. However, even in this case, the only significant independent predictor was mindware instantiation, although both numeracy and Faith in Intuition were marginally significant. While Need for Cognition and cognitive reflection emerged as predictors of latency and confidence latency detection effects, respectively, neither of the two overall regression models was significant.

Table S5. Summary of the regression analyses predicting detection effect sizes

	Latency detection effect		CON detection effect		CON LAT detection effect	
	β	*p*	*β*	*p*	*β*	*p*
Constant		.811		.944		.641
Mindware	–.02	.792	**.15**	**.008**	–.01	.919
Cognitive ability	.12	.070	.07	.279	–.07	.249
Numeracy	.01	.891	.12	.052	.07	.292
Need for Cognition	**.13**	**.021**	.01	.786	–.00	.968
Faith in Intuition	.02	.702	–.09	.061	–.05	.354
Cognitive reflection	–.05	.445	.07	.248	**.13**	**.032**
	R^2 = .02, p = .063		R^2 = .09, p < .001		R^2 = .01, p = .199	

Note. The table contains standardized regression coefficients (*β*) with their respective significance. R^2 denotes adjusted r-square for the model with appropriate significance. CON: confidence, LAT: latency. Significant regression coefficients are presented in bold.

Lastly, we have rerun the final regression model from the main manuscript where standard individual difference predictors, mindware instantiation, and conflict detection indices predicted overall reasoning accuracy but again used detection effects instead of detection efficiencies. As can be seen from Table S6, three detection effects explained 2% of additional variance in conflict reasoning accuracy over and above standard individual difference predictors and mindware instantiation. However, only the confidence detection effect emerged as a significant independent predictor, consistently with the results presented in the main manuscript.

Table S6. Summary of the regression analysis predicting the composite of correctly answered conflict reasoning problems

	β	*p*
Step 1		
Constant		< .001
Cognitive ability	**.15**	**.004**
Numeracy	**.27**	**.001**
Need for Cognition	.03	.524
Faith in Intuition	**–.15**	**< .001**
Cognitive reflection	**.26**	**< .001**
R^2 = .34, $F(5,378) = 40.98$, $p < .001$		
Step 2		
Constant		.690
Cognitive ability	**.11**	**.026**
Numeracy	**.23**	**< .001**
Need for Cognition	.01	.763
Faith in Intuition	**–.13**	**.001**
Cognitive reflection	**.21**	**< .001**
Mindware	**.25**	**< .001**
ΔR^2 = .05, $F(1,377) = 31.67$, $p < .001$		
Step 3		
Constant		.685
Cognitive ability	**.10**	**.047**
Numeracy	**.22**	**< .001**
Need for Cognition	.01	.837
Faith in Intuition	**–.12**	**.003**
Cognitive reflection	**.20**	**< .001**
Mindware	**.23**	**< .001**
Latency detection effect	.02	.674
CON detection effect	**.14**	**.002**
CON LAT detection effect	–.02	.712
ΔR^2 = .02, $F(3,374) = 3.55$, $p = .015$		

Note. The table contains standardized regression coefficients (*β*) with their respective significance. R^2 and ΔR^2 denote adjusted r-square for the initial model and change in r-square at the 2nd and 3rd step of the regression with appropriate change statistics. LAT: latency, CON: confidence. Significant regression coefficients are presented in bold.

Section E

Correlations between detection effects observed on individual problems and indices

In the main manuscript, we present correlations between detection efficiency indices calculated on the basis of different reasoning problems and conflict detection indices. From these analyses, it is evident that detection efficiency rather domain specific, i.e. detection efficiencies are mostly uncorrelated across reasoning problems and detection indices. For completeness, we here present correlations between detection effects (average latency increase or confidence decrease on the conflict in comparison with no-conflict reasoning problems) observed on every individual reasoning problem and detection index (Table S7). The results again suggest low domain generality of detection ability, detection effects in individual problems and indices are mostly uncorrelated with those from other problems and indices (see also Frey & De Neys, 2017). While there are only 14 significant correlations, all are in the expected direction with the exception of one negative correlation between latency increase on base-rate neglect problems and confidence decrease in conjunction fallacy items. Significant positive correlations suggest at least some consistency in successful detection on different detection indices for given reasoning problem.

Table S7. Correlations between detection effects observed on individual reasoning problems and detection indices

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Belief bias latency	1	212	225	247	270	212	225	247	270	212	225	247
2. Bat-and-ball latency	**.16**	1	219	242	212	273	219	242	212	273	219	242
3. Base-rate neglect latency	.10	.08	1	278	225	219	294	278	225	219	294	278
4. Conjunction fallacy latency	**.18**	−.00	**.15**	1	247	242	278	341	247	242	278	341
5. Syllogism CON	**.24**	.09	.02	.08	1	212	225	247	270	212	225	247
6. Bat-and-ball CON	.13	**.20**	.02	−.05	−.11	1	219	242	212	273	219	242
7. Base-rate neglect CON	.11	.05	**.26**	.07	.07	.07	1	278	225	219	294	278
8. Conjunction fallacy CON	−.02	.00	**−.14**	**.12**	−.08	.05	.07	1	247	242	278	341
9. Syllogism CON LAT	**.12**	−.12	.04	−.04	**.18**	−.02	**.17**	−.08	1	212	225	247
10. Bat-and-ball CON LAT	−.01	.05	−.01	−.06	−.03	**.36**	.02	−.03	.05	1	219	242
11. Base-rate neglect CON LAT	−.00	.04	.03	−.06	.02	.04	**.13**	−.00	−.01	−.00	1	278
12. Conjunction fallacy CON LAT	.06	.11	−.10	.01	.05	.04	−.04	.05	.02	**.19**	.05	1

Note. Correlations between detection effects for given reasoning problem and detection index are given below the diagonal. Numbers above the diagonal show the number of observations on which the correlation for given pair of variables was based. Correlations that appear in bold are significant at $p < .05$. CON: confidence, LAT: latency.

Section F

Research problems used in the present study

Syllogistic reasoning problems

Set A	Set B
All flowers need light Roses are flowers Roses need light (No-conflict: Valid/Believable)	All flowers need light Roses need light Roses are flowers (Conflict: Invalid/Believable)
All things made of wood can be used as fuel Trees can be used as fuel Trees are made of wood (Conflict: Invalid/Believable)	All things made of wood can be used as fuel Trees are made of wood Trees can be used as fuel (No-conflict: Valid/Believable)
All mammals can walk Spiders can walk Spiders are mammals (No-conflict: Invalid/Unbelievable)	All mammals can walk Whales are mammals Whales can walk (Conflict: Valid/Unbelievable)
All vehicles have wheels Boats are vehicles Boats have wheels (Conflict: Valid/Unbelievable)	All vehicles have wheels Trolley suitcases have wheels Trolley suitcases are vehicles (No-conflict: Invalid/Unbelievable)
All birds have wings Crows are birds Crows have wings (No-conflict: Valid/Believable)	All birds have wings Crows have wings Crows are birds (Conflict: Invalid/Believable)
All cannons fire bullets Water cannons are cannons Water cannons fire bullets (Conflict: Valid/Unbelievable)	All cannons fire bullets Guns fire bullets Guns are cannons (No-conflict: Invalid/Unbelievable)
All flowering plants have leafs Fern has leafs Fern is a flowering plant (No-conflict: Invalid/Unbelievable)	All flowering plants have leafs Cacti are flowering plants Cacti have leafs (Conflict: Valid/Unbelievable)

All dogs have snouts	All dogs have snouts
Labradors have snouts	Labradors are dogs
Labradors are dogs	Labradors have snouts
(Conflict: Invalid/Believable)	(No-conflict: Valid/Believable)

Neutral syllogistic reasoning tasks (same in both sets):

All X are Y	All X are Y
All Y are Z	All Z are Y
All X are Z	All Z are X
(Valid)	(Invalid)

Base-rate neglect problems

Karl: (Conflict version)
In a study 1000 people were tested. Among the participants there were 5 twenty-year olds and 995 sixty-year-olds. Karl is a randomly chosen participant of the study.

Karl likes to listen to rock music and goes to concerts. He painted his cell phone in the colors of his favorite football team.

Which statement is most likely?

Karl is twenty (0)
Karl is sixty (1)

Karl: (No-conflict version)
In a study 1000 people were tested. Among the participants there were 995 twenty-year olds and 5 sixty-year olds. Karl is a randomly chosen participant of the study.

Karl likes to listen to rock music and goes to concerts. He painted his cell phone in the colors of his favorite football team.

Which statement is most likely?

Karl is twenty (1)
Karl is sixty (0)

Paul: (Conflict version)
In a study 1000 people were tested. Among the participants there were 5 people who drive a used Honda and 995 people who drive a BMW.

Paul is a randomly chosen participant of the study. Paul is 38. He works in a steel plant. He lives in a small apartment in the outskirts of Detroit. His wife has left him.

Which statement is most likely?

Paul drives a used Honda (0)

Paul drives a BMW (1)

Paul: (No-conflict version)
In a study 1000 people were tested. Among the participants there were 995 people who drive a used Honda and 5 people who drive a BMW. Paul is a randomly chosen participant of the study.

Paul is 38. He works in a steel plant. He lives in a small apartment in the outskirts of Detroit. His wife has left him.

Which statement is most likely?

Paul drives a used Honda (1)

Paul drives a BMW (0)

Stan: (Conflict version)
In a study 1000 people were tested. Among the participants there were 5 dentists and 995 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.

Which statement is most likely?

Stan is a dentist (0)

Stan is a rock singer (1)

Stan: (No-conflict version)
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.

Which statement is most likely?

Stan is a dentist (1)

Stan is a rock singer (0)

Jo: (Conflict version)
In a study 1000 people were tested. Among the participants there were 5 whose favorite television series is Star Trek and and 995 whose favorite is America's Next Top Model. Jo is a randomly chosen participant of the study.

Jo is 26 years old and is doing graduate studies in physics. He stays at home most of the time and likes to play video games.

Which statement is most likely?

Jo's favorite series is Star Trek (0)

Jo's favorite series is America's Next Top Model (1)

Jo: (No-conflict version)
In a study 1000 people were tested. Among the participants there were 995 whose favorite television series is Star Trek and and 5 whose favorite is America's Next Top Model. Jo is a randomly chosen participant of the study.

Jo is 26 years old and is doing graduate studies in physics. He stays at home most of the time and likes to play video games.

Which statement is most likely?

Jo's favorite series is Star Trek (1)

Jo's favorite series is America's Next Top Model (0)

Jack: (Conflict version)
In a study 1000 people were tested. Among the participants there were 995 lawyers and 5 engineers. Jack is randomly chosen participant of this study.

Jack is 36 years old. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and writing computer programs.

Which statement is most likely?

Jack is an engineer (0)

Jack is a lawyer (1)

Jack: (No-conflict version)
In a study 1000 people were tested. Among the participants there were 5 lawyers and 995 engineers. Jack is randomly chosen participant of this study.

Jack is 36 years old. He is not married and is somewhat introverted. He likes to spend his free time reading science fiction and writing computer programs.

Which statement is most likely?

Jack is a lawyer (0)

Jack is an engineer (1)

Kai: (Conflict version)
In a study 1000 people were tested. Among the participants there were 995 nurses and 5 doctors. Kai is a randomly chosen participant of this study.

Kai is 34 years old and lives in a beautiful home in a fancy suburb. Kai is well-spoken and very interested in politics. A lot of Kai's time is spent in career development.

Which statement is most likely?

Kai is a doctor (0)

Kai is a nurse (1)

Kai: (No-conflict version)
In a study 1000 people were tested. Among the participants there were 5 nurses and 995 doctors. Kai is a randomly chosen participant of this study.

Kai is 34 years old and lives in a beautiful home in a fancy suburb. Kai is well-spoken and very interested in politics. A lot of Kai's time is spent in career development.

Which statement is most likely?

Kai is a nurse (0)

Kai is a doctor (1)

Kurt: (Conflict version)
In a study 1000 people were tested. Among the participants there were 995 who live in a farmhouse and 5 who live in a condo. Kurt is a randomly chosen participant of this study.

Kurt works on Wall Street and is single. He works long hours and wears Armani suits to work. He likes wearing shades.

Which statement is most likely?

Kurt lives in a condo (0)

Kurt lives in a farmhouse (1)

Kurt: (No-conflict version)
In a study 1000 people were tested. Among the participants there were 5 who live in a farmhouse and 995 who live in a condo. Kurt is a randomly chosen participant of this study.

Kurt works on Wall Street and is single. He works long hours and wears Armani suits to work. He likes wearing shades.

Which statement is most likely?

Kurt lives in a farmhouse (0)

Kurt lives in a condo (1)

Lilly: (Conflict version)
In a study 1000 people were tested. Among the participants there were 995 executive managers and 5 kindergarten teachers. Lilly is a randomly chosen participant of this study.

Lilly is 37 years old. She is married and has 3 kids. Her husband is a veterinarian. She is committed to her family and always watches the daily cartoon shows with her kids.

Which statement is most likely?

Lilly is an executive manager (1)

Lilly is a kindergarten teacher (0)

Lilly (No-conflict version)
In a study 1000 people were tested. Among the participants there were 5 executive managers and 995 kindergarten teachers. Lilly is a randomly chosen participant of this study.

Lilly is 37 years old. She is married and has 3 kids. Her husband is a veterinarian. She is committed to her family and always watches the daily cartoon shows with her kids.

Which statement is most likely?

Lilly is a kindergarten teacher (1)

Lilly is an executive manager (0)

Neutral base-rate neglect problems (same in both sets):

In a study 1000 people were tested. Among the participants there were 5 who campaigned for George W. Bush and 995 who campaigned for John Kerry. Jim is a randomly chosen participant of this study.

Jim is 5 feet and 8 inches tall, has black hair, and is the father of two young girls. He drives a yellow van that is completely covered with posters.

Which statement is most likely?

Jim campaigned for George W. Bush (0)

Jim campaigned for John Kerry (1)

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 a randomly chosen participant of this study.

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.

Which statement is most likely?

Tom plays the trumpet (1)

Tom plays the saxophone (0)

Conjunction fallacy problems

Jeremy: (Conflict version)
Jeremy is 16. He wears old-fashioned clothes. He studies very well and is the teacher's pet. He doesn't have many friends and doesn't do well with girls.

Which statement is most likely?

Jeremy often goes to parties (1)

Jeremy often goes to parties and gets bullied (0)

Jeremy: (No-conflict version)
Jeremy is 16. He wears old-fashioned clothes. He studies very well and is the teacher's pet. He doesn't have many friends and doesn't do well with girls.

Which statement is most likely?

Jeremy often gets bullied (1)

Jeremy often gets bullied and goes to parties (0)

Jamal: (Conflict version)
Jamal is 21 and lives near Brooklyn. He has dreadlocks and drives a convertible. He is 6' 7" (6 feet and 7 inches tall) and very athletic.

Which statement is most likely?

Jamal is a gymnast (1)

Jamal is a gymnast and a basketball player (0)

Jamal: (No-conflict version)
Jamal is 21 and lives near Brooklyn. He has dreadlocks and drives a convertible. He is 6' 7" (6 feet and 7 inches tall) and very athletic.

Which statement is most likely?

Jamal is a basketball player (1)

Jamal is a basketball player and a gymnast (0)

Ellen: (Conflict version)
Ellen likes to listen to hip hop and rap music. She enjoys wearing tight shirts and jeans. She's fond of dancing and has a small nose piercing.

Which statement is most likely?

Ellen is fifty years old (1)

Ellen is fifty years old and a dj in her spare time (0)

Ellen: (No-conflict version)
Ellen likes to listen to hip hop and rap music. She enjoys wearing tight shirts and jeans. She's fond of dancing and has a small nose piercing.

Which statement is most likely?

Ellen is a dj in her spare time (1)

Ellen is dj in her spare time and is fifty years old (0)

James: (Conflict version)
James is 26. He lives in Manhattan. He likes to wear designer clothes and acts somewhat stuck-up. On Sunday he plays golf with his father.

Which statement is most likely?

James volunteers in the day care center in his free time (1)

James volunteers in the day care center in his free time and works as a stock broker (0)

James: (No-conflict version)

James is 26. He lives in Manhattan. He likes to wear designer clothes and acts somewhat stuck-up. On Sunday he plays golf with his father.

Which statement is most likely?

James works as a stock broker (1)

James works as a stock broker and volunteers in the day care center in his free time (0)

Jake: (Conflict version)
Jake is 20. He grew up in a poor family in a neglected neighborhood. He is quite violent and already served a short sentence in prison.

Which statement is most likely?

Jake plays the violin (1)

Jake plays the violin and is jobless (0)

Jake: (No-conflict version)
Jake is 20. He grew up in a poor family in a neglected neighborhood. He is quite violent and already served a short sentence in prison.

Which statement is most likely?

Jake is jobless (1)

Jake is jobless and plays the violin (0)

Jon: (Conflict version)
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 statement is most likely?

Jon plays in a rock band (1)

Jon plays in a rock band and is an accountant (0)

Jon: (No-conflict version)
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 statement is most likely?

Jon is an accountant (1)

Jon is an accountant and plays in a rock band (0)

Lisa: (Conflict version)
Lisa is in her twenties and jobless. She applied for two different part-time jobs. For the dress shop job, there are 7 other applicants, and for the job in the variety store, there is only 1 other applicant.

Which statement is most likely?

Lisa will be offered the job in the dress shop (1)

Lisa will be offered the job in the dress shop and the job in the variety store (0)

Lisa: (No-conflict version)
Lisa is in her twenties and jobless. She applied for two different part-time jobs. For the dress shop job, there are 7 other applicants, and for the job in the variety store, there is only 1 other applicant.

Which statement is most likely?

Lisa will be offered the job in the variety store (1)

Lisa will be offered the job in the variety store and the job in the dress shop (0)

Jay: (Conflict version)
Jay is a 29-year-old male. He has served a short time in prison. He has been living on his own for 2 years now. He has an older car and listens to punk music.

Which statement is most likely?

Jay works as a lawyer (1)

Jay works as a lawyer and has a tattoo (0)

Jay: (No-conflict version)
Jay is a 29-year-old male. He has served a short time in prison. He has been living on his own for 2 years now. He has an older car and listens to punk music.

Which statement is most likely?

Jay has a tattoo (1)

Jay has a tattoo and works as a lawyer (0)

Neutral conjunction fallacy problems (same in both sets):

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 statement is most likely?

The car with its lights on is black (1)
The car with its lights on is black and is a Volkswagen

In a music store there are different kinds of instruments, including 20 stringed instruments. 15 of the instruments are guitars. 5 of the instruments are violins. One of the instruments is damaged.

Which statement is most likely?

The damaged instrument is stringed (1)

The damaged instrument is stringed and is a guitar (0)

Bat-and-ball problems

Apple & orange: (Conflict version)
An apple and an orange weigh 160 grams altogether. The apple weighs 100 grams more than the orange. How much does the orange weigh?

60 grams (0)

30 grams (1)

Apple & orange: (No-conflict version)
An apple and an orange weigh 160 grams altogether. The apple weighs 100 grams. How much does the orange weigh?

30 grams (0)

60 grams (1)

PCs & MACs: (Conflict version)
In a computer shop there are 250 PCs and MACs altogether. There are 200 more PCs than MACs. How many MACs are there in the shop?

50 MACs (0)

25 MACs (1)

PCs & MACs: (No-conflict version)
In a shop there are 250 PCs and MACs altogether. There are 200 PCs. How many MACs are there in the shop?

25 MACs (0)

50 MACs (1)

Plumber & electrician: (Conflict version)
In total, a plumber and an electrician work 240 days. The electrician works 200 days more than the plumber. How many days does the plumber work?

40 days (0)

20 days (1)

Plumber & electrician: (No-conflict version)
In total, a plumber and an electrician work 240 days. The electrician works 200 days. How many days does the plumber work?

20 days (0)

40 days (1)

Book & magazine: (Conflict version)
Altogether, a book and a magazine have 330 pages. The book has 300 pages more than the magazine. How many pages does the magazine have?

30 pages (0)

15 pages (1)

Book & magazine: (No-conflict version)
Altogether, a book and a magazine have 330 pages. The book has 300 pages. How many pages does the magazine have?

15 pages (0)

30 pages (1)

Pencil & eraser: (Conflict version)
A pencil and an eraser cost $1.10 in total. The pencil costs $1 more than the eraser. How much does the eraser cost?

10 cents (0)

5 cents (1)

Pencil & eraser: (No-conflict version)
A pencil and an eraser cost $1.10 in total. The pencil costs $1. How much does the eraser cost?

10 cents (1)

5 cents (0)

Cheese & bread: (Conflict version)
A cheese and a bread cost $2.90 in total. The cheese costs $2 more than the bread. How much does the bread cost?

90 cents (0)

45 cents (1)

Cheese & bread: (No-conflict version)
A cheese and a bread cost $2.90 in total. The cheese costs $2. How much does the bread cost?

90 cents (1)

45 cents (0)

Sandwich & soda: (Conflict version)
A sandwich and a soda cost $2.50 in total. The sandwich costs $2 more than the soda. How much does the soda cost?

50 cents (0)

25 cents (1)

Sandwich & soda: (No-conflict version)
A sandwich and a soda cost $2.50 in total. The sandwich costs $2. How much does the soda cost?

50 cents (1)

25 cents (0)

Coffee & cookie: (Conflict version)
A coffee and a cookie cost $2.40 in total. The coffee costs $2 more than the cookie. How much does the cookie cost?

40 cents (0)

20 cents (1)

Coffee & cookie: (No-conflict version)
A coffee and a cookie cost $2.40 in total. The coffee costs $2. How much does the cookie cost?

40 cents (1)

20 cents (0)

Neutral bat-and-ball problems (same in both sets):

A magazine costs $3 and a drink costs $2. How much do they cost together?

$4 (0)

$5 (1)

An elephant weighs 6 tons and a car weighs 3 tons. How much do they weight together?

9 tons (1)

8 tons (0)

Section G

Analyses for two numeracy measures presented separately

In all analyses pertaining to individual difference predictors of conflict detection and conflict problem accuracy in the main manuscript, we used a composite index based on two numeracy measures as an index of participants' numerical abilities. This was done mainly to simplify the analyses as the two numeracy measures were moderately mutually correlated (r = .25) and to counteract the potential problems of the low reliability of our objective numeracy measure, the Berlin numeracy test (α = .41). However, as the second numeracy measure that we used in our research – Subjective numeracy scale – consisted of self-report items, its inclusion in a single composite score along with the objective numeracy test may have arguably confounded the interpretation of our results as indicating participants' numerical abilities, rather than their perception of their numerical abilities. Therefore, for completeness, we here present all analyses pertaining to individual predictors of conflict detection and conflict problem accuracy for the two numeracy measures separately (see tables S8, S9, and S10).

As can be seen from both the correlational and regression results, both the Berlin numeracy test and Subjective numeracy scale exhibit the same patterns of relationships with other variables in the study. While the correlations are at some points somewhat stronger in the case of Subjective numeracy scale, this likely merely reflects higher reliability of the latter measure in comparison with the objective numeracy test. Crucially, the regression results presented here are completely consistent with the key conclusions reported in the main manuscript. Namely, while neither of the two numeracy measures significantly predicts conflict detection efficiency, both objective numeracy and subjective numeracy indicator are significant predictors of the accuracy on conflict reasoning problems after taking into account all other standard individual difference predictors, mindware instantiation, and conflict detection efficiency.

Table S8. Correlations between latency and confidence detection efficiency, conflict reasoning accuracy composite, standard individual difference predictors, and mindware instantiation

	1.	2.	3.	4.	5.	6.	7.	8.	9.
1. Detection efficiency: latency	1								
2. Detection efficiency: confidence	**.24**	1							

3. Conflict reasoning accuracy	.07	**.27**	1						
4. Cognitive reflection	.07	**.22**	**.47**	1					
5. Faith in Intuition	.00	–.03	**–.26**	**–.12**	1				
6. Need for Cognition	**.14**	**.10**	**.24**	**.22**	–.09	1			
7. Berlin numeracy test	**.10**	**.14**	**.34**	**.34**	**–.12**	**.22**	1		
8. Subjective numeracy scale	.09	**.17**	**.42**	**.33**	–.07	**.36**	**.25**	1	
9. Cognitive ability	**.12**	**.19**	**.46**	**.50**	**–.19**	**.25**	**.39**	**.40**	1
10. Mindware	.04	**.22**	**.49**	**.35**	**–.15**	**.21**	**.17**	**.33**	**.35**

Note. Correlations pertaining to detection efficiencies are based on 384 observations, others on 399 observations. Correlations that appear in bold are significant at $p < .05$.

Table S9. Summary of the regression analysis predicting latency and confidence detection efficiency

	Detection efficiency: latency		Detection efficiency: confidence	
	β	p	β	p
Constant		.001		.651
Mindware	–.02	.771	**.14**	**.010**
Cognitive ability	.07	.251	.05	.444
Berlin numeracy test	.05	.357	.05	.358
Subjective numeracy sc.	.02	.741	.05	.376
Need for Cognition	**.11**	**.046**	.01	.865
Faith in Intuition	.03	.540	.02	.652
Cognitive reflection	–.00	.981	**.12**	**.048**
	$R^2 = .01$, $F(7,376) = 1.69$, $p = .109$		$R^2 = .07$, $F(7,376) = 4.85$, $p < .001$	

Note. The table contains standardized regression coefficients (β) with their respective significance. R^2 denotes adjusted r-square for the model with appropriate F-statistics. Significant regression coefficients are presented in bold.

Table S10. Summary of the regression analysis predicting the composite of correctly answered conflict reasoning problems

	β	p
Step 1		
Constant		.231
Cognitive ability	**.15**	**.004**
Berlin numeracy test	**.10**	**.037**
Subjective numeracy scale	**.23**	**< .001**
Need for Cognition	.03	.513
Faith in Intuition	**–.16**	**< .001**
Cognitive reflection	**.26**	**< .001**
$R^2 = .34$, $F(6,377) = 34.00$, $p < .001$		

Step 2		
Constant		**.047**
Cognitive ability	**.11**	**.031**
Berlin numeracy test	**.11**	**.016**
Subjective numeracy scale	**.18**	**< .001**
Need for Cognition	.01	.750
Faith in Intuition	**–.14**	**.001**
Cognitive reflection	**.20**	**< .001**
Mindware	**.25**	**< .001**
ΔR^2 = .05, $F(1,376) = 32.01$, $p < .001$		
Step 3		
Constant		.051
Cognitive ability	**.11**	**.037**
Berlin numeracy test	**.10**	**.020**
Subjective numeracy scale	**.18**	**< .001**
Need for Cognition	.02	.731
Faith in Intuition	**–.14**	**.001**
Cognitive reflection	**.19**	**< .001**
Mindware	**.24**	**< .001**
Detection efficiency: LAT	–.02	.642
Detection efficiency: CON	**.11**	**.008**
ΔR^2 = .01, $F(2,374) = 3.52$, $p = .031$		

Note. The table contains standardized regression coefficients (β) with their respective significance. R^2 and ΔR^2 denote adjusted r-square for the initial model and change in r-square at the 2nd and 3rd step of the regression with appropriate change statistics. LAT: latency, CON: confidence. Significant regression coefficients are presented in bold.

Supplementary references

Frey, D., & De Neys, W. (2017). Is Conflict Detection in Reasoning Domain General ? In *Proceedings of the Annual Meeting of the Cognitive Science Society, 39* (pp. 391–396).

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *The Quarterly Journal of Experimental Psychology*, *71*(5), 1188–1208.

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.