# Conflict Detection across Various Probabilistic Reasoning Tasks

Rowan Haen[1], Eva Janssen[1], Peter Verkoeijen[2,3], Wim de Neys[4], & Tamara van Gog[1]

[1] Department of Education, Utrecht University, The Netherlands

[2] Brain and Learning Research Group, Expertise Centre Future-proof Education, Learning and Innovation Centre, Avans University of Applied Sciences, The Netherlands

[3] Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam, The Netherlands

[4] CNRS, University of Paris, France

## Author Notes

E-mail and ORCID:

Rowan Haen: r.haen@uu.nl, ORCID: 0009-0009-0220-553X

Eva Janssen: e.m.janssen@uu.nl, ORCID: 0000-0001-9064-8473

Peter Verkoeijen: ppjl.verkoeijen@avans.nl, ORCID: 0000-0002-8085-5038

Wim De Neys: wim.de-neys@parisdescartes.fr, ORCID: 0000-0003-0917-8852

Tamara van Gog: t.vangog@uu.nl, ORCID: 0000-0003-3766-6255

Correspondence concerning this manuscript should be addressed to Rowan Haen, Department of Education, Utrecht University, P.O. Box 80140, 3508TC Utrecht, The Netherlands, E: r.haen@uu.nl

**Abstract**

Many biased reasoners seem sensitive to the conflict between their heuristic, biased responses and logical/probabilistic principles, known as conflict detection (CD). As CD has been found to predict receptiveness to feedback and training, it seems educationally relevant. However, to date, CD has mostly been studied with classic "heuristics-and-biases tasks", and it is unclear how consistent it is within participants across tasks that share a mindware component. Therefore, we explored whether CD is: 1) found in more complex (Bayesian) reasoning tasks, and 2) consistent across probabilistic reasoning tasks of varying complexity that share a mindware component (i.e., base-rate, inverse-fallacy, and Bayesian-inference-tasks). Results showed that the proportion conflict detectors decreased with increasing task complexity. Moreover, for the Bayesian-inference-task, CD-presence depended on analysis method (types of errors included). We found little evidence for consistency in CD.

*Keywords*: Dual process theory, Probabilistic reasoning, Conflict detection, Confidence, Judgement accuracy

**Conflict Detection across Various Probabilistic Reasoning Tasks**

Decades of research on reasoning and decision-making have demonstrated that even well-educated individuals frequently deviate from fundamental principles of logic and mathematics (e.g., Kahneman, 2011). This deviation often arises from reliance on intuitive rules-of-thumb, commonly known as "heuristics" (Tversky & Kahneman, 1974). While heuristics can yield valid conclusions in many instances, they may also lead to reasoning biases in situations where they conflict with the laws of logic/probability (De Neys & Bonnefon, 2013).

Traditionally, biased reasoning was often attributed to a failure to recognize the conflict between heuristic responses and logical/probabilistic principles (Evans & Stanovich, 2013; Kahneman, 2011). Interestingly, however, more recent research has shown that many biased reasoners do show some sensitivity to this conflict: They have less confidence in their heuristic responses on items that contain such a conflict than on items that do not (i.e., where the heuristic response is the logically/probabilistically correct response). This is known as conflict detection (CD).

CD seems meaningful for education, as individuals who exhibit CD are more receptive to feedback (Janssen et al., 2020) and benefit more from training (e.g., Boissin et al., 2022; Franiatte et al., 2024). In theory, CD can thus be used as a predictor of training success. However, most CD-studies have used classic "heuristic-and-biases tasks", involving basic logical/probabilistic principles (De Neys, 2015), whereas daily-life reasoning tasks are often more complex, requiring more knowledge and skills (i.e., "mindware"; Perkins, 1995). Therefore, it is important to 1) determine if CD is also found in more complex reasoning tasks. In research on syllogisms, it has been hypothesized that there may be a link between complexity and CD but findings are mixed (Brisson et al., 2018; Trippas et al., 2017). Moreover, previous studies with tasks requiring different types of mindware (e.g., logic vs.

probability) concluded that CD is not consistent across individuals (Frey & de Neys, 2017; Srol & de Neys, 2021). It would be relevant to 2) know to what extent CD is consistent across tasks that share a mindware component.

We address these two questions by using three probabilistic reasoning tasks (i.e., base-rate, inverse-fallacy, and Bayesian-inference-tasks) that share a mindware component (base-rate importance), but differ in complexity (i.e., extent to which additional mindware components are required for accurate performance). Regarding the first question, we hypothesize that CD will occur on all task types, but that as task complexity increases, the proportion of conflict detectors within the group of biased reasoners decreases (because cues pointing to base-rate importance are less salient and fewer people possess the necessary mindware; e.g., Stanovich, 2018). Regarding the second question, we did not formulate a hypothesis given the lack of prior research; rather, we exploratively analyze the consistency of CD across these three tasks and the relationship between task performance and CD/CD-effect size.

## Method

### Participants

We recruited 100 Dutch-speaking bachelor's and master's students (≥18 years) via Prolific (www.prolific.ac), who received £5 (ca. €5.75) for their participation. After excluding three participants who failed one or more attention checks (e.g., "This is an attention check, please select answer A") and one who answered all items within a few seconds (i.e., did not read the questions), the final sample consisted of 96 participants.

### Design, Materials, and Procedure

This study was approved by the faculty ethics review board of the first author's institute. All materials were presented in Qualtrics survey software. The study had a within-subjects design. After providing consent, participants were randomly assigned to one of three

item lists, in which all participants were presented with three blocks of tasks: 4 Bayesian-inference-tasks, 8 base-rate-tasks, and 4 inverse-fallacy-tasks, half in conflict version, where the heuristic and normative answer conflicted, and half in no-conflict version, where the heuristic and normative response were the same. The item lists were created from among 24 pairs of conflict and no-conflict versions of items, in such a way that participants never got both versions of the same item. The order of the tasks and items was fixed[1] (see Supplementary Materials, Table S1).

***Base-rate-tasks (BR)***

Participants completed eight base-rate-tasks adapted from De Neys & Glumicic (2008). For example:

> In a study 1000 people were tested. Among the participants there were 995 nurses and 5 doctors. Paul is a randomly chosen participant of this study. Paul is 34 years old. He lives in a beautiful home in a posh suburb. He is well spoken and very interested in politics. He invests a lot of time in his career.
>
> What is most likely?
>
> A. Paul is a doctor
>
> B. Paul is a nurse

The heuristic response cued by the description would be that Paul is most likely a doctor. However, given the base-rate, the normative response[2] is that Paul is most likely a nurse. No-conflict versions were constructed by switching the base-rate (995 doctors) so that the heuristic and normative response aligned.

[1] NB: Participants were also asked to explain their answers on the last item of each task type (always a conflict item) and to complete the AOT-13 (Stanovich & Toplak, 2023), as pilot-tests for another study.

[2] The normative status of the base-rate response is sometimes debated. Given our item design with extreme base-rates and moderate descriptions, this problem is minimized and even a very approximate Bayesian reasoner should pick the base-rate response (see Supplementary Materials, Section A).

***Inverse-fallacy-tasks (IF)***

Participants completed four newly developed inverse-fallacy-tasks (Koehler, 1996). For example:

> You have been experiencing various symptoms for a week, but you do not know what you have. Before you consult your doctor, you decide to look on the internet. You look up the symptoms you have, and you come across a specific condition. Your symptoms match perfectly, and it turns out that as many as 90% of people who have this condition have the same symptoms you have. You conclude that there is a good chance that you also have this condition, since you have these symptoms.
>
> A. Conclusion is correct
>
> B. Conclusion is incorrect

The heuristic response would be that the conclusion is correct, as many people intuitively interpret this inversely (having these symptoms = 90% chance of having this condition) and fail to recognize that no base rate is mentioned. Without base-rate information, the conclusion cannot be justified based on the available information and, therefore, the conclusion is incorrect[3]. No-conflict versions were constructed by reversing the conditional statement so that the heuristic and normative response are the same: Your symptoms match perfectly, and it turns out that as many as 90% of people who have these symptoms do have this specific condition.

***Bayesian-inference-tasks (BI)***

Participants completed four Bayesian-inference-tasks, adapted from Eddy (1982). For example:

[3] One might argue that, in theory, the conclusion could be correct (e.g., if the base rate is high), and that, therefore, answer option B cannot be considered normative. However, the core of the question is not: What should you have concluded? (i.e., cannot be determined), but rather: Is the conclusion that was drawn correct? The answer to that question is 'no'.

In a certain city, 10% of the population has a rare disease. A diagnostic test has been developed to detect this disease. The test correctly identifies whether a person has the disease or not in 80% of the cases and is wrong 20% of the time. A randomly selected person from this city is tested, and the test results come back positive. What do you think are the chances that this person has the disease?

A) 81% - 100%

B) 61% - 80%

C) 41% - 60%

D) 21% - 40%

E) 0% - 20%

The most popular heuristic response is 80% (Villejoubert & Mandel, 2002), reflecting the test's hit-rate (i.e., "hit-rate-heuristic"), but the normative answer, based on Bayes' Theorem is 30.77%[4]. In all items we used, the hit-rate was always above 75% and the prior probability below 10%. No-conflict versions were constructed by changing the base-rate to 50-50 and ensuring the hit-rate and the false positive rate sum to one, aligning the heuristic and normative response.

Note that the complexity of these three types of tasks differs. Base-rate-tasks are the least complex; base-rate-information is explicitly mentioned in the task, recognizing its importance (which can be done intuitively) is sufficient for solving the problem correctly. Inverse-fallacy-tasks are more complex; recognition of base-rate importance is again required for solving the problem correctly, but there are no explicit base-rate cues mentioned in this task, which may make it harder to detect conflict and rely on intuition alone. Moreover, knowledge about principle $P(H|D) \neq P(D|H)$ is needed to solve the problem correctly. Bayesian-inference-tasks are the most complex; even though the task does mention base-rate

[4] Bayes' theorem: $P(H|D) = P(D|H) * P(H) / P(D)$. Solution: $0.8 * 0.1 / 0.26 = 30.77\%$.

information explicitly, its relevance is less clear (for those without the right mindware. Moreover, simply recognizing its importance (which possibly can be intuitive) is not sufficient for solving the problem correctly. To do so, one also needs specific mindware on Bayesian reasoning, which must be applied through deliberate thought**.**

### *Confidence*

After each item, participants rated their confidence in their answers using a slider to choose any value from 0% (*not at all confident*) to 100% (*completely confident*).

### Analyses

Analyses were conducted with RStudio 4.1.2 (R Core Team, 2023). There were no outliers on the confidence ratings (cf. Šrol & De Neys, 2021). Due to a technical error in Qualtrics, we could not use data from one conflict and one no-conflict item of the Bayesian-inference-task. To test for conflict detection, we conducted linear mixed-effects models with confidence as outcome measure. The within-subjects factor item version (conflict/no-conflict) was entered as fixed effect, and participant and item number were entered as random effects. As conflict detection concerns the within-subject difference in mean confidence when incorrectly answering conflict items vs correctly answering no-conflict items, we only included confidence on incorrect conflict trials and correct no-conflict trials. The difference in mean confidence represents the CD effect-size. Participants who did not answer any conflict items incorrectly or any no-conflict items correctly were excluded from this analysis (i.e., it is not possible to contrast the confidence levels on incorrect conflict and correct no-conflict items within participants unless at least one no-conflict item is answered correctly and one conflict item incorrectly).

## Results

Data and R-code available on OSF: https://osf.io/x6u93/?view_only=f3761aca54b24d23b5bf76d70ac254c7.

**Reasoning Accuracy**

Table 1 presents participants' average reasoning accuracy. Consistent with previous research (e.g., Pennycook et al., 2015), participants performed significantly better on no-conflict than on conflict versions of the base-rate-tasks ($t = 7.28$, $p < .001$), inverse-fallacy-tasks ($t = 9.67$, $p < .001$) and Bayesian-inference-tasks ($t = 9.80$, $p < .001$). As expected, base-rate-tasks were the easiest, followed by inverse-fallacy-tasks and Bayesian-inference-tasks. Performance on base-rate conflict items was higher than expected based on other studies (e.g., 20% in Frey et al., 2018; 39% in Janssen et al., 2021). Because inverse-fallacy-tasks were developed for this study, we cannot compare accuracy with other studies. Performance on Bayesian-inference-tasks seem to align with previous research (4% in McDowell & Jacobs, 2017, with an open-ended answer format). Additionally, we conducted exploratory correlation analyses on accuracy for conflict items. Only base-rate-tasks and inverse-fallacy-tasks were weakly correlated ($r = .28$, $p < .05$; see Supplementary Materials, Section B/Table S2).

**Table 1**

*Average Accuracy (SD) on the Reasoning Tasks in Whole Sample (N = 96)*

| | **Conflict** | **No-conflict** |
|---|---|---|
| **Task:** | | |
| BR | 77.4% (31.3) | 95.2% (14.1) |
| IF | 34.9% (36.4) | 87.5% (21.8) |
| BI | 9.9% (21.9) | 70.5% (39.2) |

**Conflict Detection**

In the CD-analyses biased reasoners who solved at least one no-conflict item correctly and at least one conflict item incorrectly per task type were included. However, for Bayesian-inference-tasks, we made a distinction between different kinds of errors, which we did not

initially planned. Unlike typical “heuristic-and-biases tasks”, Bayesian-inference-tasks do not have a dichotomous answer format, allowing for different response strategies to result in different types of errors (Hafenbrädl & Hoffrage, 2015). We realized that providing erroneous responses other than the hit-rate can imply two things: either one applies a different incorrect heuristic, or one does not use a heuristic but lacks the knowledge to solve the problem. In the latter case, this might indicate one is not biased by intuitive heuristics. Therefore, we report the CD-analyses both with all errors and with “hit-rate-errors" only.

Table 2 shows the number of biased reasoners per task type, and their average confidence ratings on correctly solved no-conflict items and incorrectly solved conflict items (for group-level averages as a function of response accuracy, see Supplementary Materials, Table S3). Results showed that participants were less confident about their performance on incorrectly solved conflict items compared to correctly solved no-conflict items on base-rate-tasks, $\beta = 0.08$, $SE = 0.08$, $t(405.75) = 9.15$, $p < .001$; inverse-fallacy-tasks, $\beta = 0.12$, $SE = 0.04$, $t(213.50) = 2.65$, $p = .009$, and Bayesian-inference-tasks, including all errors, $\beta = 0.04$, $SE = 0.04$, $t(172.68) = 3.29$, $p = .001$; however, this difference was not significant when including only hit-rate-errors on Bayesian-inference-tasks, $\beta = 0.04$, $SE = 0.04$, $t(152.1) = 1.01$, $p = .31$.

***Individual Level***

Following Frey et al. (2018), we divided biased reasoners into three subgroups: those with lower confidence ratings on incorrect conflict problems compared to correct no-conflict problems (i.e., *subgrou*p *detection*), those with the same confidence ratings for both problem types (i.e., *subgroup same*), and those with higher confidence on incorrect conflict problems compared to correct no-conflict problems (i.e., *subgroup reverse*) and analysed the percentage of conflict detectors within the group of biased reasoners, per task type (Table 2). Consistent with our hypothesis, the pattern in the descriptive statistics suggest that the percentage of

conflict detectors within the biased group seems to decrease as task complexity increased from base-rate to inverse-fallacy to Bayesian-inference-tasks.

**Table 2**

*Conflict Detection: Confidence Ratings in (Sub)Groups of Biased Reasoners, per Task*

| | **Whole biased group** | **Subgroup detection** | **Subgroup same** | **Subgroup reverse** |
|---|---|---|---|---|
| **BR** | **(n = 40)** | **(n = 28)** | **(n = 1)** | **(n = 11)** |
| % of biased reasoners | 100% | 70% | 2.5% | 27.5% |
| % of entire sample | 41.7% | 29.2% | 1.0% | 11.5% |
| *M* confidence (*SD*): | | | | |
| no-conflict correct | 84.2% (14.4) | 89.4% (11.3) | 56.5% | 73.6% (13.6) |
| conflict incorrect | 66.4% (20.7) | 62.1% (21.9) | 56.5% | 78.3% (12.3) |
| *M* CD-effect size (*SD*) | -17.8 (22.4) | -27.3 (24.7) | - | - |
| **IF** | **(n = 81)** | **(n = 45)** | **(n = 9)** | **(n = 27)** |
| % of biased reasoners | 100% | 55.6% | 11.1% | 33.3% |
| % of entire sample | 84.4% | 46.9% | 9.4% | 28.1% |
| *M* confidence (*SD*): | | | | |
| no-conflict correct | 82.8% (14.4) | 82.8% (13.8) | 98.1% (5.7) | 77.8% (14.3) |
| conflict incorrect | 78.2% (17.4) | 69.2% (15.6) | 98.1% (5.7) | 86.5% (12.9) |
| *M* CD-effect size (*SD*) | -4.7 (13.9) | -13.6 (20.8) | - | - |
| **BI** (All errors) | **(n = 74)** | **(n = 36)** | **(n = 14)** | **(n = 24)** |
| % of biased reasoners | 100% | 48.6% | 18.9% | 32.4% |
| % of entire sample | 77.1% | 37.5% | 14.5% | 25.0% |

| | | | | |
|---|---|---|---|---|
| *M* confidence (*SD*): | | | | |
| no-conflict correct | 82.6% (17.2) | 84.6% (15.9) | 95% (11.1) | 72.3% (16.5) |
| conflict incorrect | 76.7% (20.8) | 65.9% (21.3) | 95% (11.1) | 82.3% (14) |
| *M* CD-effect size (*SD*) | -5.9 (18.6) | -18.7 (26.6) | - | - |
| **BI** (Hit-rate errors) | **(n = 65)** | **(n = 27)** | **(n = 14)** | **(n = 24)** |
| % of biased reasoners | 100% | 41.5% | 33.1% | 36.9% |
| % of entire sample | 67.7% | 28.1% | 14.5% | 25.0% |
| *M* confidence (*SD*): | | | | |
| no-conflict correct | 82.9% (16.9) | 86% (14.6) | 95% (11.1) | 72.3% (16.5) |
| conflict incorrect | 80.9% (17.5) | 72.5% (18.3) | 95% (11.1) | 82.3% (14) |
| *M* CD-effect size (*SD*) | -2.0 (24.4) | -13.6 (23.4) | - | - |

Note. Confidence in non-hit-rate errors (BI): *M* = 53.13% (*SD* = 21.76).

**Consistency**

To determine consistency in CD, we used cross-tables (cf. Janssen et al., 2021) to count how many participants showed bias, and how many of those showed CD, across two or three tasks (Table 3). Most biased reasoners were inconsistent detectors, showing CD on only one or two tasks.

**Table 3**

*Consistency of CD across Tasks*

| | **Number of biased reasoners** | **Consistent detectors** | **Inconsistent detectors** | **Non-detectors** |
|---|---|---|---|---|
| **Comparison between tasks:** | | | | |
| BI–IF | 37 | 12 (32.43%) | 23 (62.16%) | 2 (5.41%) |
| BR–BI | 32 | 11 (34.38%) | 16 (50.00%) | 5 (15.63%) |

| | | | | |
|---|---|---|---|---|
| IF–BI | 64 | 14 (21.88%) | 36 (56.25%) | 14 (21.88%) |
| BR–IF–BI | 30 | 6 (20.00%) | 24 (80.00%) | 0 (0.00%) |
| BR–BI (H) | 26 | 7 (26.92%) | 14 (53.85%) | 5 (19.23%) |
| IF–BI (H) | 55 | 11 (20.00%) | 30 (54.54%) | 14 (25.45%) |
| BR–IF–BI (H) | 24 | 5 (20.83%) | 19 (79.17%) | 0 (0.00%) |

Note. (H) = Hit-rate errors only.

Additionally, as our tasks differed in complexity but shared a mindware component, one might expect that individuals' performance on simpler tasks is associated with showing (stronger) CD on more complex tasks. We tested this in two ways. First, we correlated CD-effect size and performance on conflict items across task types (see Supplementary Materials, Table S4). None of these correlations were significant. Whereas previous studies found a positive *within-task* correlation (e.g., Mevel et al., 2015; Pennycook et al., 2015); here, it was only significant for base-rate-tasks.

Second, we used logistic regression to test whether performance on simpler conflict tasks would predict CD-presence in more complex tasks (see Supplementary Materials, Table S5). No significant relationships were found. We did find a significant relationship between performance on Bayesian-inference-tasks (with all and hit-rate errors) and CD-presence on inverse-fallacy-tasks. Given that we did not expect this relationship and considering the number of conducted tests, this result should be interpreted with caution.

## Discussion

This study investigated whether conflict detection (CD) appears across probabilistic reasoning tasks of varying complexity and whether it is consistently demonstrated by individuals across these tasks.

### Task Complexity and Conflict Detection

At the group level, our findings replicate and extend findings from prior research. We found that participants were, on average, less confident about their incorrect responses on conflict tasks than their correct responses on no-conflict tasks. This CD-effect was not only found in base-rate-tasks as in previous research (e.g., De Neys & Glumicic, 2008), but also on the more complex-inverse-fallacy tasks, and shows that participants sensed that something was wrong with their biased response.

Importantly, however, regarding the even more complex Bayesian-inference-task, whether CD was found depended on the analysis choices. That is, unlike most tasks used in CD-research that have a dichotomous answer format, the Bayesian-inference-task had a multiple-choice format, allowing for other heuristic or non-heuristic erroneous responses. Hence, some of our participants might not have been biased by the hit-rate-heuristsic, yet might not have known how to solve the problem. Therefore, we conducted the CD-analyses both with and without non-hit-rate errors. Intriguingly, we found evidence of CD when including all errors, but not when analysing only hit-rate errors. Possibly, (some) individuals who make other errors deviate from the hit-rate heuristic response because they realize it is incorrect, but also lack the mindware to come up with the correct answer, and thus exhibit lower confidence (see Table 2). On dichotomous tasks, it is hard to identify such individuals, as they are likely to opt for the correct answer when they realize it is not the heuristic answer, even if they are unable to reason through why it is correct. However, when only looking at individuals biased by the hit-rate heuristic, we have to conclude that at the group level, we did not find evidence of CD on the Bayesian-inference-task.

At the individual level, the numerical data revealed a pattern showing that, consistent with our hypothesis, the percentage of conflict detectors seems to decrease with increasing task complexity, suggesting that fewer participants were sensitive to their errors on the more complex tasks. Presumably, fewer individuals possess the necessary mindware for CD to arise

on the more complex tasks (on which realizing the relevance of base-rates is necessary but not sufficient to avoid a heuristic answer). Moreover, the mean CD-effect size displayed by conflict detectors also seemed lower on the more complex inverse-fallacy and Bayesian-inference-tasks than on the base-rate-tasks.

**Consistency in Conflict Detection**

Prior research found little evidence for consistency in CD across tasks (Frey & de Neys, 2017; Srol & de Neys, 2021), but there were substantial differences in mindware required on those tasks. We examined consistency across tasks that shared an important knowledge component (i.e., base-rate importance). However, of the biased reasoners, only 20% showed CD across all three tasks. Most participants were inconsistent detectors, showing CD on only one or two tasks. Moreover, we found no relationship between CD/CD-effect size on a complex task and performance on a simpler task.

**Are Cues Key to CD?**

A possible explanation for the reduced CD-effect (size) in more complex tasks, may lie in the saliency and/or perceived relevance of cues in the tasks. Explicit cues, such as (extreme) base-rates, have been shown to influence both performance accuracy and CD-effect size on base-rate-tasks (Pennycook et al., 2015). In contrast, base-rate cues in inverse-fallacy-tasks are more implicit and not salient (have to be inferred). In Bayesian-inference-tasks, while the base rate is as salient as in base-rate-tasks, its perceived relevance may be less apparent (Bar-Hillel, 1980). Consequently, when cues are less salient or perceived as less relevant, this may lead to poorer performance, fewer individuals displaying CD, and a smaller CD-effect among conflict detectors.

Moreover, in the Bayesian-inference-tasks, cues that trigger heuristic responses –such as the hit-rate of a diagnostic test– might be particularly compelling because individuals are frequently exposed to predictive tests in real-life context (e.g., COVID-tests, weather

forecasting). This repeated exposure, often without correction, might reinforce the intuitive but incorrect interpretation of a test's hit-rate as the positive predictive value. As a result, the heuristic intuition becomes especially strong, reducing the likelihood for CD to occur, as CD is caused by the activation difference between "logical" and "heuristic" intuitions (De Neys, 2022). This may explain both the smaller group of conflict detectors and the smaller CD-effect for those who do detect. Furthermore, the varying salience and relevance of cues across tasks could also explain the lack of consistency in CD, despite the shared knowledge component. Proper activation of this mindware component may not occur in tasks where logical cues are less salient or relevant, or where heuristic cues are particularly strong.

**Limitations and Future Research**

A limitation of this study is the lack of randomization in task order, which may have introduced order effects that influenced participants' performance. Specifically, the high accuracy observed on the base-rate-tasks may, at least in part, be a consequence of their position in the task sequence. Completing the more complex Bayesian-inference-tasks first may have primed participants to reason more analytically, which constrains the interpretation of the results regarding the consistency between base-rate-tasks and other tasks. To more rigorously assess consistency across tasks, future studies should employ counterbalanced task orders.

Another limitation is that the findings regarding the pattern of decreasing percentages of conflict detectors with increasing task complexity should be interpreted with caution as we did not statistically test it, due to the methodological complexity (i.e., partially overlapping subgroups of participants) that precludes straightforward comparison of groups. Future research should aim to confirm this pattern by employing alternative study designs and/or larger sample sizes.

Concerning the Bayesian-inference-task, we used a multiple-choice format rather than a dichotomous one, which may have affected the results. For more direct comparisons between tasks, especially with respect to the shared base-rate relevance component, future studies may benefit from using a uniform response format across tasks (e.g., higher vs lower than 50% in Bayesian inference tasks). Note, however, that differentiating between types of errors, which provided useful information in the present study, is not possible when using a dichotomous answer format. Moreover, the difference in findings regarding CD-presence when including all vs. only hit-rate-errors, which significantly impacted the results, shows the importance of considering a) the analysis approach in future research (we recommend reporting both analyses), and b) what this means for the interpretation of accuracy and CD-data on tasks with dichotomous answer options.

Future research should aim to replicate these results with probabilistic as well as logical reasoning tasks, as they hold significant implications for the design and implementation of educational interventions. That is, our findings suggest that CD has predictive value only for training effects on the same tasks and not necessarily for more complex tasks that share a knowledge component (i.e., transfer; which is in line with training studies suggesting that transfer is hard to establish; e.g., Heijltjes et al., 2014; Van Peppen et al., 2022). Since we observed CD also in more complex tasks and training research mostly focused on simpler tasks, future training research should determine whether these findings extend to complex tasks, where conflict detectors may benefit more from the training (Boissin et al., 2022; Franiatte et al., 2024).

**Conclusion**

This study suggests that complexity is a boundary condition for the CD-effect: the more complex the tasks (i.e., mindware required), the lower the performance accuracy, the number of conflict detectors, and the CD-effect size.

**Acknowledgment**

This research was funded by the Netherlands Initiative for Education Research (NRO, project number 40.5.21945.325).

**Declaration of interest**

None.

## References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgements. *Acta Psychologica, 44*(3), 211-233. https://doi.org/10.1016/0001-6918(80)90046-3

Boissin, E., Caparos, S., Voudouri, A., & De Neys, W. (2022). Debiasing System 1: Training favours logical over stereotypical intuiting. *Judgment and Decision Making, 17(4),* 646-690. https://doi.org/10.1017/S1930297500008895

Brisson, J., Schaeken, W., Markovits, H., & De Neys, W. (2018). Conflict detection and logical complexity. *Psychologica Belgica*, *58*(1), 318. https://doi.org/10.5334/pb.448

De Neys, W. (2015). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science, 10*(6), 558-575. https://doi.org/10.1177/1745691611429354

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. https://doi.org/10.1016/j.tics.2013.02.001

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

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge, U.K.: Cambridge University Press

Evans, J. & Stanovich, K. E. (2013). Dual-Process Theories of Higher Cognition: Advancing the Debate. Perspectives on Psychological Science, 8*(3),* 223-241. https://doi.org/10.1177/1745691612460685

Franiatte, N., Boissin, E., Delmas, A., & De Neys, W. (2024). Boosting debiasing: Impact of repeated training on reasoning. *Learning and Instruction, 89*, 101845. https://doi.org/10.1016/j.learninstruc.2023.101845

Frey, D., & De Neys, W. (2017). Is conflict detection in reasoning domain general? *Proceedings of the Annual Meeting of the Cognitive Science Society, 39*, 391-396.

Frey, D., Johnson, E. D., & De Neys, W. (2018). Individual differences in conflict detection during reasoning. *Quarterly Journal of Experimental Psychology, 71(5)*, 1188-1208. https://doi.org/10.1080/13546783.2019.1708793

Hafenbrädl, S., & Hoffrafe, U. (2015). Toward an ecological analysis of Bayesian inferences: how task characteristics influence responses. *Frontiers of Psychology, 6*, 939. https://doi.org/10.3389/fpsyg.2015.00939

Heijltjes, A., Van Gog, T., Leppink, J., & Paas, F. (2014). Improving critical thinking: Effects of dispositions and instructions on economics students' reasoning skills. *Learning and Instruction*, *29*, 31-42. https://doi.org/10.1016/j.learninstruc.2013.07.003

Janssen, E. M., Raoelison, M., & de Neys, W. (2020). "You're wrong!": The impact of accuracy feedback on the bat-and-ball problem. *Acta Psychologica*, *206,* Article 103042. https://doi.org/10.1016/j.actpsy.2020.103042

Janssen, E. M., Velinga, S. B., de Neys, W., & Van Gog, T. (2021). Recognizing biased reasoning: Conflict detection during decision-making and decision-evaluation. *Acta Psychologica, 217*, 103322. https://doi.org/10.1016/j.actpsy.2021.103322

Kahneman, D. (2011). Thinking, Fast and Slow. Farrar, Straus and Giroux.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*(1), 1-53. https://doi.org/10.1017/S0140525X00041157

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin, 143(12)*, 1273. http://dx.doi.org/10.1037/bul0000126

Mevel, K., Poirel, N., Rossi, S., Cassotti, M., Simon, G., Houde, O., & De Neys, W. (2015). Bias detection: Response confidence evidence for conflict sensitivity in the ratio bias task. *Journal of Cognitive Psychology, 27(2),* 227-237. https://doi.org/10.1080/20445911.2014.986487

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. https://doi.org/10.1016/j.cogpsych.2015.05.001

Van Peppen, L. M., van Gog, T., Verkoeijen, P. P., & Alexander, P. A. (2022). Identifying obstacles to transfer of critical thinking skills. *Journal of Cognitive Psychology*, *34*(2), 261-288. https://doi.org/10.1080/20445911.2021.1990302

Perkins, D. (1995). *Outsmarting IQ: The emerging science of learnable intelligence*. Free Press.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Šrol, J., & De Neys, W. (2021). Predicting individual differences in conflict detection and bias susceptibility during reasoning. *Thinking & Reasoning, 27(1),* 38-68. https://doi.org/10.1080/13546783.2019.1708793

Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24(4),* 423-444. https://doi.org/10.1080/13546783.2018.1459314

Stanovich, K. E., & Toplak, M. E. (2023). Actively open-minded thinking and its measurement. *Journal of Intelligence, 11*(2), 27. https://doi.org/10.3390/jintelligence11020027

Trippas, D., Thompson, V. A., & Handley, S. J. (2017). When fast logic meets slow belief: Evidence for a parallel-processing model of belief bias. *Memory & cognition, 45*(4), 539. https://doi. org/10.3758/s13421-016-0680-1

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Villejoubert, G. & Mandel, D. R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory & Cognition, 30*(2), 171-178. https://doi.org/10.3758/BF03195278

**Supplementary Materials**

**Conflict Detection across Various Probabilistic Reasoning Tasks**

---

**Table S1**

*Fixed Task Order of each List (C = conflict, NC = no-conflict)*

| | List 1 | List 2 | List 3 |
|---|---|---|---|
| **BI** | 1-C | 6-C | 3-C |
| | 2-NC | 5-NC | 4-NC |
| | 3-NC | 1-NC | 6-NC |
| | 4-C | 2-C | 5-C |
| **BR** | 1-C | 2-C | 6-C |
| | 2-NC | 1-NC | 5-NC |
| | 3-C | 4-C | 7-C |
| | 4-NC | 3-NC | 8-NC |
| | 5-C | 9-C | 10-C |
| | 6-NC | 10-NC | 9-NC |
| | 7-NC | 11-NC | 12-NC |
| | 8-C | 12-C | 11-C |
| **IF** | 1-C | 6-C | 3-C |
| | 2-NC | 5-NC | 4-NC |
| | 3-NC | 1-NC | 6-NC |
| | 4-C | 2-C | 5-C |

**Section A: Justification base-rate-task**

Critics of the base-rate-task (e.g., Gigerenzer, Hell, & Blank, 1988; see also Barbey & Sloman, 2007) have noted that when reasoners take a Bayesian approach and combine base-rate probabilities with stereotypical descriptions, it can lead to complications if the description is highly diagnostic. For instance, consider an example where males and females are the two groups, and the description says that Person ‘A’ is ‘pregnant.’ In such a case, one would always conclude that Person ‘A’ is a woman, regardless of the base rates. More moderate descriptions, like ‘kind’ or ‘funny,’ help avoid this issue (see Boissin et al., 2023). Moreover, extreme base-rates (e.g., 997/3, 996/4, 995/5) further ensure that even an approximate Bayesian reasoner would select the response cued by the base-rates (see De Neys, 2014). In this study, we used items from De Neys and Glumicic (2008), who systematically tested stereotypical descriptions, ensuring that only moderately diagnostic descriptions were employed throughout and used extreme base-rates.

**Table S2**

*Correlations of Accuracy on Conflict Items between Task Types (N = 96)*

| Task type: | **BR** | **IF** |
|---|---|---|
| **BR** | - | - |
| **IF** | .28* | - |
| **BI** | .06 | .06 |
| **BI(H)** | -.06 | -.06 |

* = $p < .05$. (H) = Hit-rate errors only.

**Section B: Exploratory Analyses**

We also conducted correlation analyses on accuracy for conflict items (Table S5). Given that the different tasks partly rely on the same mindware, one would expect significant positive correlations in performance on conflict items, which we found between the base-rate and inverse-fallacy-tasks, but not with the Bayesian-inference-task (possibly due to the very low average accuracy). We also explored whether performance on the base-rate or inverse-fallacy-tasks would predict making other errors than the hit-rate error on the Bayesian-inference-task, because recognizing the importance of the base-rate (a necessary but not sufficient condition for performing well on the Bayesian-inference-task), might mean that participants do not fall for the hit-heuristic answer option on conflict items, even if they lack the skills to come up with the correct answer. However, this was not the case; base-rate: $r(94) = -0.04$, $p = .702$; inverse-fallacy: $r(93) = -0.115$, $p = .270$.

**Table S3**

*Group-Level Averages (SD) on Confidence Ratings as Function of Response Accuracy*

| | **Conflict: correct** | **Conflict: incorrect** | **No-conflict: correct** | **No-conflict: incorrect** |
|---|---|---|---|---|
| **BR** | | | | |
| Participants per group | 88 | 40 | 96 | 10 |
| Average confidence (%) | 85.9 (17.2) | 66.4 (20.7) | 89.4 (13.1) | 77.4 (18.5) |
| **IF** | | | | |
| Participants per group | 52 | 81 | 96 | 24 |
| Average confidence (%) | 75.8 (18.1) | 78.2 (17.4) | 82.6 (14.3) | 77.9 (17.7) |
| **BI (All errors)** | | | | |
| Participants per group | 14 | 93 | 75 | 36 |
| Average confidence (%) | 64.8 (23.9) | 72.3 (23) | 82.4 (17.1) | 66.1 (22.1) |
| **BI (Hit-rate errors)** | | | | |
| Participants per group | 12 | 93 | 75 | 36 |
| Average confidence (%) | 64.8 (23.9) | 78.9 (19.7) | 82.4 (17.1) | 66.1 (22.1) |

**Table S4**

*Correlations between Performance on Conflict Items and CD-Effect Size Within and Across Task Types*

| | CD-effect size: | | | |
|---|---|---|---|---|
| | **BR** | **IF** | **BI** | **BI (H)** |
| Performance on Conflict items: | | | | |
| **BR** | .39*[40] | -.19[81] | .07[74] | .11[65] |
| **IF** | .25[40] | .14[81] | -.10[74] | -.05[65] |
| **BI** | -.18[40] | -.03[81] | -.15 [74] | - |
| **BI (H)** | -.18[40] | -.03[81] | - | -.26*[65] |

* = $p < .05$. (H) = Hit-rate errors only. Superscripts indicate *n*.

**Table S5**

*Logistic regressions Models Predicting CD-presence from Accuracy on Conflict Items across Task Types*

| **CD-presence** | **Accuracy on conflict items** | **n** | **B** | **SE** | **z** | **p** |
|---|---|---|---|---|---|---|
| IF | BR | 40 | -.90 | .70 | -1.27 | .20 |
| BI | BR | 40 | .15 | .71 | .22 | .83 |
| BI (H) | BR | 40 | .51 | .80 | .64 | .52 |
| BR | IF | 40 | 1.70 | 1.34 | 1.27 | .20 |
| BI | IF | 81 | -.29 | .67 | -.40 | .66 |
| BI (H) | IF | 81 | .27 | .71 | .38 | .71 |
| BR | BI | 74 | 0.67 | 1.46 | 0.48 | .65 |
| BR | BI (H) | 65 | .70 | 1.50 | .46 | .65 |
| IF | BI | 74 | 3.12 | 1.50 | 2.10 | .04* |
| IF | BI (H) | 65 | 3.10 | 1.50 | 2.08 | .04* |

* = $p < .05$. (H) = Hit-rate errors only.