

Supplemental analyses: Confidence

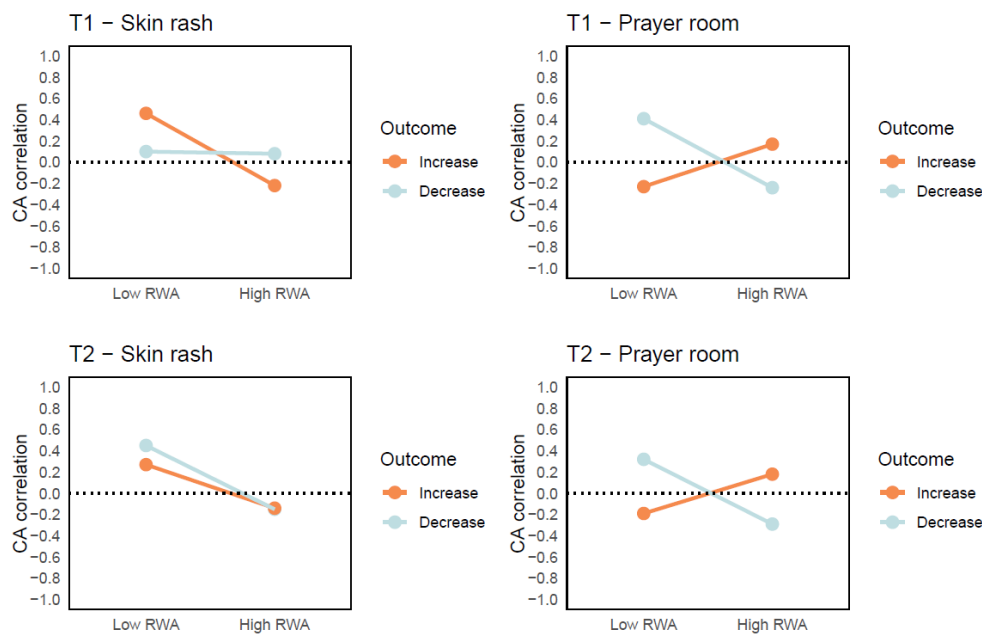
Experiment 1

We expected motivated tendencies to lead participants to express greater confidence in belief-consistent answers, regardless of the accuracy of these responses (H4).

Results

We examined H4, that participants in the polarized scenario would be more confident when their responses aligned with their ideological beliefs (i.e., RWA) than when responses countered these beliefs, regardless of response accuracy. Thus, we expected high-RWA participants to be more confident when they answered that more generous rules led to an *increase* in extremism support, whether this answer was correct or not, and vice versa for low-RWA participants. No effects of ideology on confidence in responses were expected in the neutral scenarios. *S Confidence Figure 1* shows the result as correlation coefficients for the confidence-accuracy relationship in each condition. In line with expectations, the figure shows a positive confidence-accuracy correlation in belief-consistent conditions, and conversely, a negative correlation in belief-inconsistent conditions. That is, when presented with a scenario outcome that countered their beliefs, participants were more confident when their responses were wrong, than when they answered correctly. To test the hypothesis statistically, we used multilevel modelling with confidence as dependent variable. We started by examining a model with conclusion accuracy, RWA, scenario, outcome as predictors, including their four-way interaction (see *S Confidence Table 1.*). A significant interaction between conclusion accuracy, RWA, scenario and outcome would support the hypothesis. In line with H4, results showed that the four-way interaction was significant ($p = .003$). Breaking down this interaction, we found significant three-way interactions between conclusion accuracy, RWA and outcome for both the polarized ($p = .015$) and neutral scenario ($p = .008$). As expected, there was a two-way interaction between conclusion accuracy and outcome in the polarized scenario for participants both high ($p = .011$) and low RWA ($p < .001$). Supporting H4, when *decrease* was the correct answer, high RWA participants were more confident when incorrectly

answering “increase” than when correctly answering “decrease” ($M_{correct} = 4.71$, $M_{incorrect} = 5.43$, $p = .032$). When the correct answer was *increase*, these participants were more confident when correctly answering “increase” than incorrectly “decrease” ($M_{correct} = 5.36$, $M_{incorrect} = 4.89$, $p = .122$). Mirroring this pattern, low-RWA participants were more confident when incorrectly answering “decrease” than “increase” when the correct answer was *increase* ($M_{correct} = 4.64$, $M_{incorrect} = 5.21$, $p = .077$), and more confident when correctly answering “decrease” than “increase” when the correct answer was *decrease* ($M_{correct} = 5.42$, $M_{incorrect} = 4.00$, $p = .004$). As expected, there were no such interactions between conclusion accuracy and outcome in the neutral scenario for neither high ($p = .383$) nor low-RWA participants ($p = .278$). Next, we tested a model with a five-way interaction, including also time as predictor. Results showed that this five-way interaction was not significant ($p = .084$, see S Confidence Table 1.).



S Confidence Figure 1. Correlations between confidence and conclusion accuracy in the different conditions in Experiment 1. Leftmost column displays results for the neutral scenario (effect of skin cream on skin rash), rightmost column shows results for the polarizing scenario (effect of prayer room on support for extremism). Top row displays results when the participants were presented with the problem for the first time (“T1”), bottom row displays the second time the problem was presented, now containing the calculations needed to reach the correct conclusion. Legend (“Outcome”) displays conditions in which the correct conclusion was an increase (e.g.,

increased support for extremism) or decrease, respectively. High/low RWA indicates participants ± 1 SD above mean ($n_{rash} = 158$; $n_{prayer\ room} = 165$).

Discussion

Our results extend previous findings by demonstrating that participants express more confidence in belief-consistent responses compared to belief-inconsistent responses, even when the belief-consistent response is actually incorrect (see S Confidence Figure 1).

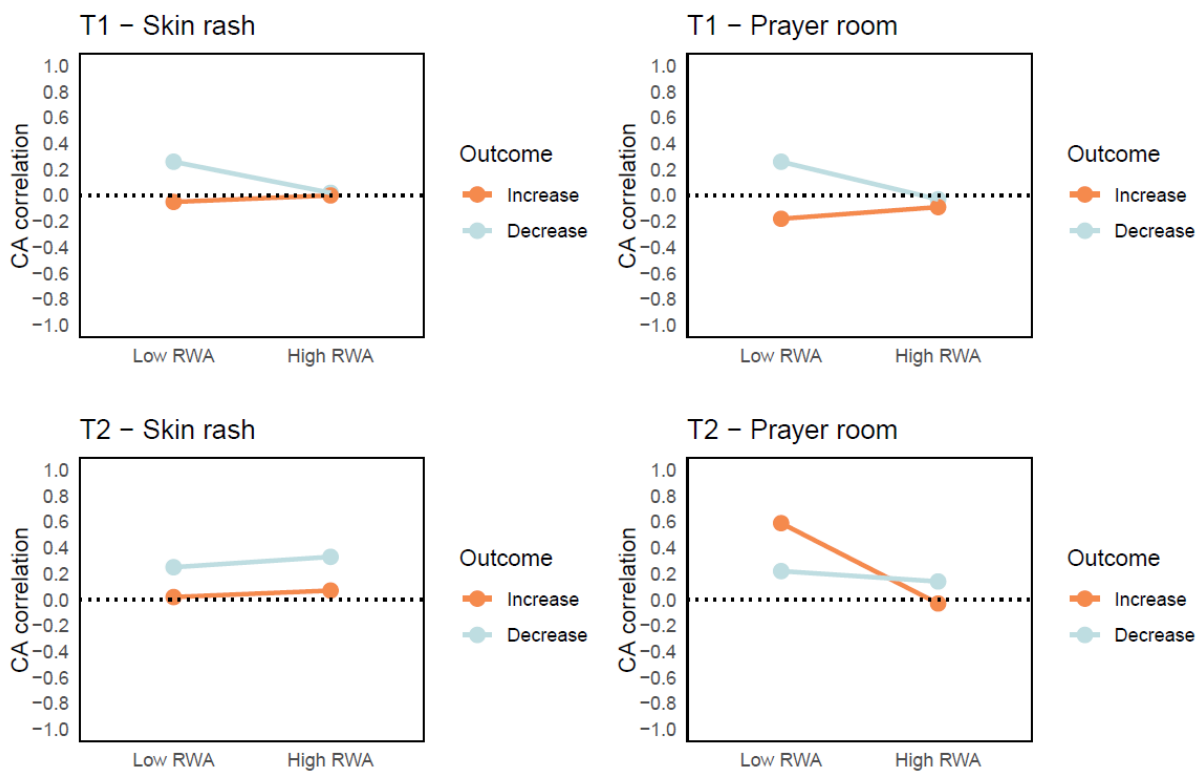
Experiment 2

As in Experiment 1, we expected that participants would be more confident in their belief-consistent responses (H4), even when this response was wrong.

Results

We examined H4, namely that participants would be more confident when responding in a belief-consistent way. Thus, we expected high-RWA participants to be more confident when they indicated that more generous rules would lead to an *increase* in support for Islamic extremism, regardless of whether this was the correct response or not (and vice versa for low-RWA participants). S Confidence Figure 2 shows this result as correlation-coefficients for the confidence-accuracy relationship in each condition. The figure suggests only a partial replication of the findings in Experiment 1; low-RWA participants indeed showed a higher positive confidence-accuracy correlation when more generous rules led to decreased rather than increased extremism at Time 1. This belief-consistent pattern was not replicated for high-RWA participants, whose confidence-accuracy relations were close to 0 regardless of outcome condition (see top right box). Importantly, the bias-consistent confidence for low-RWA participants completely disappeared, and actually inverted, at Time 2, that is, when the problem was presented with percentages (as opposed to only frequencies). For high-RWA participants, differences remained small between decrease and increase outcomes across presentations. We examined four- and

five-way interactions including conclusion accuracy, RWA, scenario, outcome, and time as predictors, and confidence as dependent variable (see S Confidence Table 2). However, results showed that neither the four-way interaction ($p = .103$) nor the five-way interaction ($p = .092$) was statistically significant.



S Confidence Figure 2. Correlations between confidence and conclusion accuracy in the different conditions in Experiment 2. Leftmost column displays results for the neutral scenario (effect of skin cream on skin rash), rightmost column shows results for the polarizing scenario (effect of prayer room on support for extremism). Top row displays results when the participants were presented with the problem for the first time (“T1”), bottom row displays the second time the problem was presented, now containing both frequencies and percentages. Legend (“Outcome”) displays conditions in which the correct conclusion was an *increase* (e.g., increased support for extremism) or *decrease*, respectively. High/low RWA indicates participants ± 1 SD above mean ($n_{rash} = 146$; $n_{prayer\ room} = 171$).

Discussion

Results regarding participants’ confidence in their responses were less clear than in Experiment 1, with no significant correlations between conclusion accuracy, RWA, scenario, outcome, and time (see S Confidence Table 2). Given the results shown in Figure 4, it appears that low-RWA participants were

more confident in belief-consistent responses (even when answering incorrectly), while high-RWA participants showed no confidence-accuracy correlation, regardless of condition. Interestingly, after being shown the tables including percentages, the more easily interpreted version of the problem, the low-RWA group's confidence-accuracy relationship switched, such that they actually became more certain when responses countered their beliefs (i.e., more generous rules for Muslim prayer rooms leading to *increased* support for Islamic extremism). Although this may suggest that people will only abandon their belief when they perceive that there is strong enough evidence in the other direction, the exploratory nature of these findings suggest a need for further research.

S Confidence Table 1. Parameter estimates (and standard error) for predictors in models of confidence in answers in Experiment 1

Predictor	Model 1	Model 2
Fixed effects		
Intercept	5.18 (0.08) ***	4.64 (0.11) ***
Accuracy	0.32 (0.07) ***	0.23 (0.07) ***
RWA	-0.05 (0.05)	-0.03 (0.05)
Scenario	-0.54 (0.08) ***	-0.55 (0.08) ***
Outcome	0.03 (0.08)	0.02 (0.08)
Time		0.40 (0.05) ***
Accuracy*RWA*Scenario*Outcome	0.32 (0.11) **	
Accuracy*RWA*Scenario*Outcome*Time		0.11 (0.06)

Note. For exact p-values, see Supplementary Table S2. n = 323, *p < .05, **p < .01, ***p < .001

S Confidence Table 2. Parameter estimates (and standard error) for predictors in models of confidence in answers in Experiment 2

Predictor	Model 1	Model 2
Fixed effects		
Intercept	5.30 (0.08) ***	4.78 (0.99) ***
Accuracy	0.42 (0.06) ***	0.23 (0.07) ***
RWA	-0.14 (0.05) **	-0.16 (0.05) **
Scenario	-0.52 (0.08) ***	-0.53 (0.08) ***
Outcome	-0.13 (0.08)	-0.13 (0.08)
Time		0.42 (0.05) ***
Accuracy*RWA*Scenario*Outcome	0.12 (0.11)	
Accuracy*RWA*Scenario*Outcome*Time		0.10 (0.06)

Note. For exact p-values, see Supplementary Table S5. n = 317, *p < .05, **p < .01, ***p < .001

