Supplemental Materials: The Political (A)Symmetry of Metacognitive Insight Into Detecting Misinformation

Michael Geers[1,2], Helen Fischer[3], Stephan Lewandowsky[4,5,6], and Stefan M. Herzog[1]

[1]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

[2]Department of Psychology, Humboldt University of Berlin, Berlin, Germany

[3]Leibniz Institut für Wissensmedien, Tübingen, Germany

[4]School of Psychological Science and Cabot Institute, University of Bristol, Bristol, United Kingdom

[5]School of Psychological Science, University of Western Australia, Perth, Australia

[6]Department of Psychology, University of Potsdam, Potsdam, Germany

Supplemental Materials: The Political (A)Symmetry of Metacognitive Insight Into Detecting Misinformation

Supplement S1: Glossary of Key Terms

Table S1

Glossary of Key Terms

Term	Definition	Application
Object-level sensitivity (d')	Ability to discriminate between two stimulus types in Signal Detection Theory (Macmillan & Creelman, 2004)	Ability to discern true from false statements
Response bias (c)	Tendency to favor one response option over the other independent of object-level sensitivity (Macmillan & Creelman, 2004)	Tendency towards responding either "true" or "false" irrespective of ground truth
Metacognitive sensitivity (meta-d')	Ability to discern one's correct from incorrect decisions based on one's confidence (Fleming & Lau, 2014)	Ability to distinguish between one's own correct and incorrect identifications of true and false statements based on one's confidence
Metacognitive efficiency (Mratio = meta-d'/d'; Mdiff = meta-$d' - d'$)	Metacognitive sensitivity relative to a given level of object-level sensitivity, operationalized either as a ratio or difference score (Fleming & Lau, 2014)	Ability to discern correct from incorrect identifications of true and false statements, controlling for discernment of true and false statements

Supplement S2: Two-Step Preregistration Procedure

Our preregistration is available on OSF (https://osf.io/veja6/). We adopted a two-step preregistration procedure: In the first step we preregistered our secondary data analysis following the template of Van den Akker et al. (2021). In the second step, we preregistered our analysis code, which we developed based on a 20% subsample of the data that served as the exploration data set. After this second step, we ran the same code again, but this time on all data. All code changes since the code preregistration (i.e., the second step of the two-step preregistration) can be seen as a git-diff on the project's GitHub repository (https://github.com/stefanherzog/misinformation_and_metacognition/compare/v0.1...v1.0). Thus, any changes in the approach of the final, second step of the preregistration do not constitute a deviation from the preregistration, but are a proper part of the overall preregistration process. Our preregistration was also informed by Srivastava's 2018 suggestions for how to aspire to the goals of preregistration in the context of research that uses more complex cognitive modeling approaches (see "Part 5: Analyses" of the first step of the preregistration for details). Notable changes between the first and second step of the preregistration include:

Knowledge Accuracy

In the first step, we only formulated metacognition analyses (i.e., meta-d', $Mratio$, and $Mdiff$), as they were the primary focus of this study. However, we conducted additional knowledge analyses (d') as part of the second preregistration step. Knowledge analyses were primarily aimed at re-establishing the main results in Garrett and Bond (2021) and to provide a more comprehensive view of the results—comprising accuracy of both knowledge and metacognition.

Statement Coding

In the first step, we planned to code statements based on participants' political party only. In the second step, we also coded statements based on participants' political

ideology, such that pro-Democrat statements were coded as congruent for liberals and pro-Republican statements were coded as congruent for conservatives (and vice versa).

In addition to coding statements based on the alignment of a statement's slant and a participant's political views (congruency), in the second step we decided to also examine statement *concordance*, which additionally considers a statement's truth. For the statement congruency coding, we would expect an effect of partisan bias on response tendency, that is, a more lenient tendency to judge statements that align with one's political views as true, and vice versa (i.e., c in SDT terms; see Batailler et al., 2022). To be able to study how political views bias information processing (i.e., facilitate or impede discrimination ability or d' in SDT terms; see also Derreumaux et al., 2023), we additionally operationalized a statement's concordance as the alignment of three factors: (i) a statement's slant, (ii) a statement's truth, and (iii) a participant's political views. This conceptualization led to three statement subgroups:

- *Concordant statements:* Pro-ingroup statements that are true and pro-outgroup statements that are false. These are concordant statements for a politically motivated person in the sense that an information-processing bias that puts more weight on information consistent with one's political views ends up putting more weight on information that points to the correct answer, thus improving performance for such statements. In short, upweighting information that points to what one wants to be true ends up improving accuracy.

- *Discordant statements:* Pro-ingroup statements that are false and pro-outgroup statements that are true. These are discordant statements for a politically motivated person in the sense that an information-processing bias that puts more weight on information consistent with one's political views ends up putting more weight on information that points to the incorrect answer, thus impairing performance for such statements. In short, upweighting information that points to what one wants to be true ends up decreasing accuracy.

- *Neutral statements:* Both true and false statements that are neither pro-ingroup nor pro-outgroup. These are neutral statements for a politically motivated person in the sense that an information-processing bias that puts more weight on information consistent with one's political views will neither improve nor impair performance for such statements, as the bias does not shift the overall weight put on information pointing either to the correct or the incorrect answer. In short, upweighting information that points to what one wants to be true ends up not affecting accuracy.

Consistent with the congruency coding, we applied this concordance coding only for participants who identified as liberal (Democratic) or conservative (Republican). For all other participants, no statement concordance score could be assigned.

Robustness Checks

As opposed to the first step of the preregistration, in the second step we decided against testing all hypotheses with different subsets of participants (e.g., participants who completed six or more waves vs. all 12 waves) and/or statements (e.g., recognized vs. non-recognized statements) to keep the number of separate models to be estimated more manageable. Instead, we estimated metacognitive efficiency across different subsets of participants or statements (see Fig. S10, Panel C). We did, however, test hypothesis H2 again separately for statements that were concordant versus discordant, with neutral statements as a benchmark, as we considered statement concordance to be of high theoretical and practical relevance.

Uninterpretable Values

Next to individual-level estimates of Mratio ≤ 0, we also defined NAs as uninterpretable values. In addition, we also considered NAs for $Mdiff$ (i.e., meta-$d' - d'$) as uninterpretable values. The model can produce NAs for $Mratio$, for instance, if $d' = 0$ or if meta-$d' > 0$ and $d' < 0$, and NAs for $Mdiff$ if no value for d' or meta-d' can be calculated. For all these values, participants were excluded from the respective analysis.

Supplement S3: Participants

Sample Demographics

See Table S2.

Dropout Across Waves

Out of the total 1,191 participants, 1,171 participants provided valid responses in wave 1 (i.e., no satisficing). Valid responses across the remaining 11 (of the total 12) waves ranged from 771 to 888 participants (see Fig. S1; for the number of completed waves across participants, see Fig. S2). Participants were allowed to return after they missed waves. Almost a third of participants (27%) completed all 12 waves, while three-quarters (75%) of the sample completed at least half of the waves (see Fig. S11 showing that our individual-level results are robust to exploring only participants who completed [i] six or more waves and [ii] all 12 waves.).

Figure S1

Number of Participants Across Waves

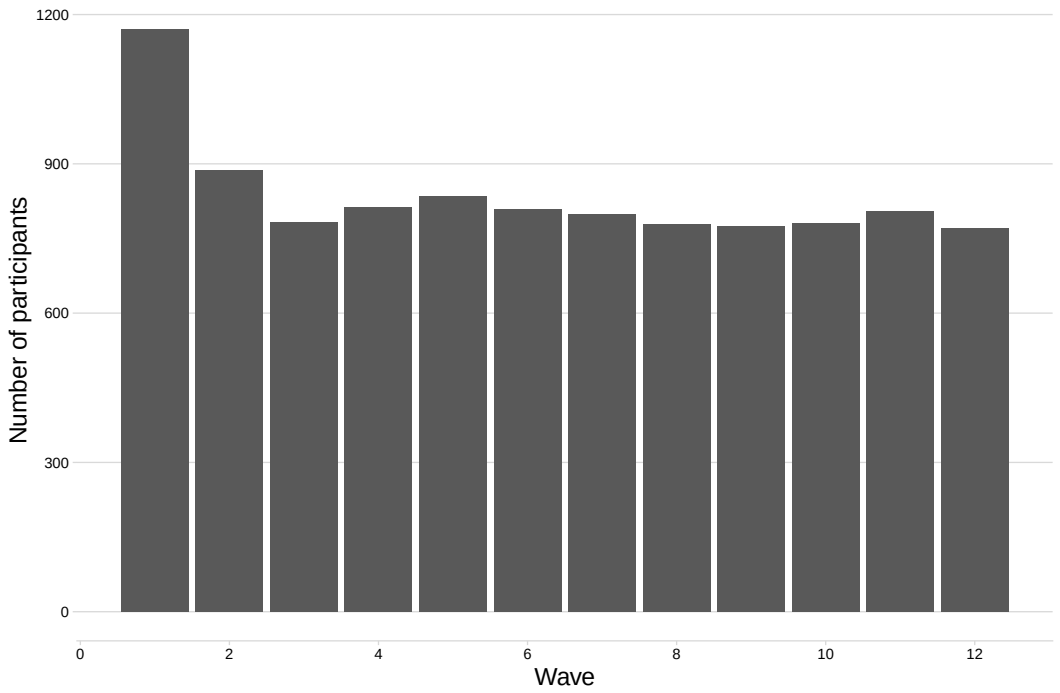


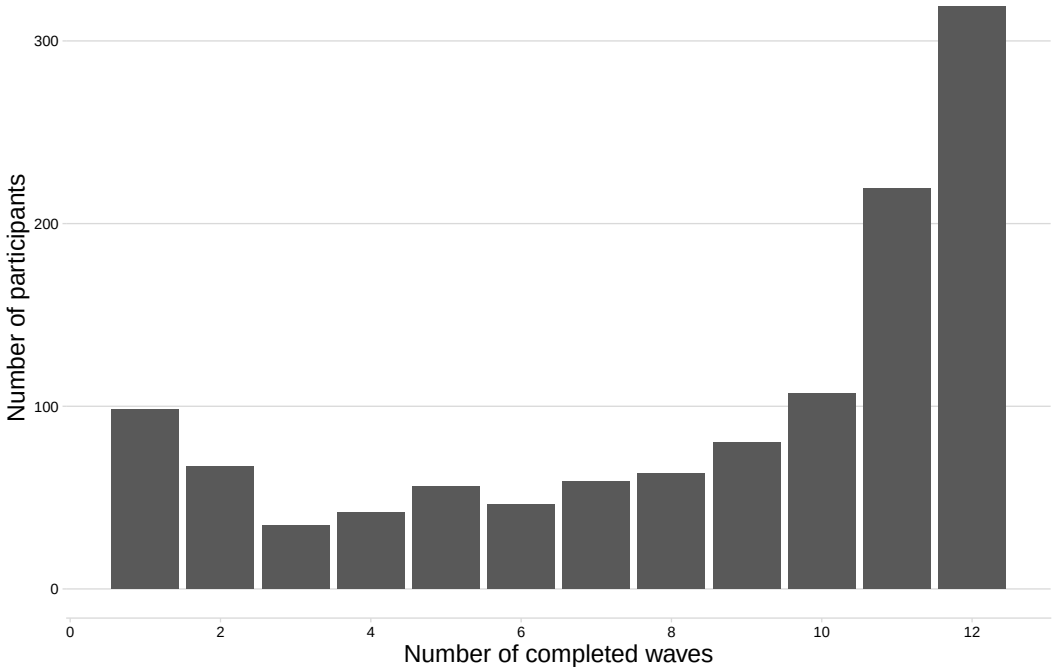
Table S2

Sample Demographics

	No.	Percent
N	1,191	100
Age		
18-29	175	14.7
30-44	285	23.9
45-59	295	24.8
60+	436	36.6
Gender		
Female	514	43.2
Male	677	56.8
Education		
No high school	45	3.8
High school graduate	354	29.7
Some college	287	24.1
2-year college degree	118	9.9
4-year college degree	244	20.5
Post-grad	143	12.0
Political party affiliation		
Democrat	404	33.9
Republican	322	27.0
Independent	336	28.2
Other	52	4.4
Not sure	77	6.5
Political ideology		
Very liberal	101	8.5
Liberal	155	13.0
Somewhat liberal	137	11.5
Moderate	319	26.8
Somewhat conservative	150	12.6
Conservative	206	17.3
Very conservative	123	10.3

Figure S2

Number of Completed Waves Across Participants



Supplement S4: Hypothesis Tests

While we broadly outlined the hypotheses tests in the first step of our two-step preregistration procedure, the code of step two implemented its analytic details (https://github.com/stefanherzog/misinformation_and_metacognition).

We focused on metacognitive efficiency (Mratio = meta-d'/d' or Mdiff = meta-$d' - d'$) in all our main analyses. When meta-d' equals d', participants are metacognitively optimal and can use all of the information available for the object-level task (evaluating political statements) when reporting confidence. When meta-d' is less than d', however, participants have a lower ability to report confidence judgments than expected based on the accuracy of their knowledge. In some cases, meta-d' may even be higher than d', when participants draw on hunches (Rausch & Zehetleitner, 2016; Scott et al., 2014), engage in deeper analysis of the stimulus information (Charles et al., 2013; Rabbitt & Vyas, 1981), or possess knowledge about other factors influencing task performance while forming their metacognitive assessments (Fleming & Daw, 2017).

Details on Bayesian Analyses

For the group-level models, we ran three Markov Chain Monte Carlo (MCMC) chains with 10,000 samples each via JAGS (see https://mcmc-jags.sourceforge.io/; before that we ran 10,000 samples for adaptation and 1,000 samples for burn-in, both not used in the analyses). For the individual level models, we ran three MCMC chains with 10,000 samples each via JAGS (see https://mcmc-jags.sourceforge.io/; before that we ran 1,000 samples for adaptation and 1,000 samples for burn-in, both not used in the analyses). We used the default priors specified by Fleming (2017) and assessed model convergence by calculating Gelman-Rubin statistics. These values indicated good convergence across chains for most group-level models (i.e., $\widehat{R} < 1.01$). Some group-level models did, however, yield an $\widehat{R} \geq 1.01$. Hence, for these group-level models as well as those with an $\widehat{R}$ right below the 1.01–threshold, we increased the number of adaptation samples from 10,000 to 30,000. This procedure led to just one group-level model with an $\widehat{R} \geq 1.01$, that is,

Republicans judging neutral statements.

Details on Software Used

Data were analyzed using R, version 4.3.1 (R Core Team, 2023).

H1.1 (metacognitive ideal hypothesis): Metacognitive efficiency for political statements is lower than the theoretically optimal metacognitive efficiency.

We compared the group-level posterior distribution (Median $Mratio$ + 95% CI) with a region of practical equivalence (ROPE; Kruschke, 2018). A ROPE defines an area that is considered equal to the target value (here: $Mratio = 1$) and then allows the researcher to compute the overlap between the posterior and the ROPE. We preregistered the ROPE to be $Mratio = 0.9–1.1$. Given that the model estimated a 99.9% probability that the average metacognitive efficiency in the study population would lie in our preregistered ROPE, we reject the hypothesis.

H1.2 (domain comparison hypothesis): Metacognitive efficiency for political statements is lower than for non-politicized domains such as biology and physics (Fischer et al., 2019).

As the 95% CI (0.88–1.16) in Fischer et al. (2019) broadly overlaps with the ROPE (0.9–1.1), we preregistered to employ the same ROPE comparison as for H1.1. Given that the model estimated a 99.9% probability that the average metacognitive efficiency in the study population would lie in our preregistered ROPE, we reject the hypothesis.

H2.1a (asymmetry hypothesis; political party identification): Metacognitive efficiency for political statements is lower for Republicans than for Democrats.

We visually inspected and compared the 95% CIs of the separate group-level posterior distributions for Democrats and Republicans. As preregistered, if the 95% CI of Republicans is credibly lower than that of Democrats (i.e., non-overlapping 95% CIs), we accept the hypothesis. In all other cases, we reject the hypothesis. Given that the 95% CIs of Republicans were not credibly lower than that of Democrats, we reject the hypothesis.

H2.1b (asymmetry hypothesis; political ideology): Metacognitive efficiency for political statements is lower for conservatives than for liberals.

We visually inspected and compared the 95% CIs of the separate group-level posterior distributions for liberals and conservatives. As preregistered, if the 95% CI of conservatives is credibly lower than that of liberals (i.e., non-overlapping 95% CIs), we accept the hypothesis. In all other cases, we reject the hypothesis. Given that the 95% CIs of conservatives were not credibly lower than that of liberals, we reject the hypothesis.

H2.2 (ideological extremity hypothesis): Metacognitive efficiency for political statements decreases strictly monotonically with political extremism on both sides of the political spectrum.

We visually inspected and compared the 95% CIs of the separate group-level posterior distributions for the seven ideological subgroups from 1 = very liberal to 7 = very conservative. As preregistered, if the 95% CIs are credibly lower with increasing political extremism (i.e., non-overlapping 95% CIs going from the midpoint 4 = moderate towards either 1 = very liberal or 7 = very conservative), we accept the hypothesis. In all other cases, we reject the hypothesis. Given that the 95% CIs were not credibly lower with increasing political extremism, we reject the hypothesis.

H2.3 (asymmetrical ideological extremity hypothesis): Metacognitive efficiency for political statements declines more strongly with political extremism for conservatives than for liberals.

We visually inspected and compared the 95% CIs of the separate group-level posterior distributions for the seven ideological subgroups from 1 = very liberal to 7 = very conservative. As preregistered, if the 95% CIs are credibly lower with increasing political extremism (i.e., non-overlapping 95% CIs going from the midpoint 4 = moderate towards either 1 = very liberal or 7 = very conservative) and the 95% CIs are credibly lower for conservatives compared to liberals of equal extremity (e.g., 5 = somewhat conservative vs. 3 = somewhat liberal), we accept the hypothesis. In all other cases, we reject the

hypothesis. Given that the 95% CIs were neither credibly lower with increasing political extremism nor credibly lower for conservatives compared to liberals of equal extremity, we reject the hypothesis.

H3.1 (faith in intuition for facts hypothesis): The stronger people's faith in their intuition for identifying facts, the lower their metacognitive efficiency for political statements.

We visually inspected the robust LOESS smooth and 95% confidence bands, and computed a Spearman correlation. Additionally, we computed a Bayes factor for the correlation. A Bayes factor greater than 1 provides evidence in favor of the alternative hypothesis, while a Bayes factor less than 1 provides evidence in favor of the null hypothesis, with larger Bayes factors indicating stronger evidence (Kass & Raftery, 1995). Our results revealed a weak positive association between participants' faith in intuition and their metacognitive efficiency ($rho = .08$, $p = .004$; $BF_{10} = 3.91$), leading us to reject the hypothesis.

H3.2 (truth is political hypothesis): The stronger people's belief that truth is political, the lower their metacognitive efficiency for political statements.

We visually inspected the robust LOESS smooth and 95% confidence bands, and computed a Spearman correlation. Additionally, we computed a Bayes factor for the correlation. A Bayes factor greater than 1 provides evidence in favor of the alternative hypothesis, while a Bayes factor less than 1 provides evidence in favor of the null hypothesis, with larger Bayes factors indicating stronger evidence (Kass & Raftery, 1995). Our results revealed a weak positive association between participants' beliefs that truth is political and their metacognitive efficiency ($rho = .10$, $p < .001$; $BF_{10} = 17.3$), leading us to reject the hypothesis.

H3.3 (need for evidence hypothesis): The higher people's need for evidence, the higher their metacognitive efficiency for political statements.

We visually inspected the robust LOESS smooth and 95% confidence bands, and computed a Spearman correlation. Additionally, we computed a Bayes factor for the correlation. A Bayes factor greater than 1 provides evidence in favor of the alternative hypothesis, while a Bayes factor less than 1 provides evidence in favor of the null hypothesis, with larger Bayes factors indicating stronger evidence (Kass & Raftery, 1995). Our results revealed no association between participants' need for evidence and their metacognitive efficiency ($rho = -.04$, $p = .23$; $BF_{10} = 0.14$), leading us to reject the hypothesis.

H3.4 (age hypothesis): Metacognitive efficiency for political statements declines with age.

We visually inspected the robust LOESS smooth and 95% confidence bands, and computed a Spearman correlation. Additionally, we computed a Bayes factor for the correlation. A Bayes factor greater than 1 provides evidence in favor of the alternative hypothesis, while a Bayes factor less than 1 provides evidence in favor of the null hypothesis, with larger Bayes factors indicating stronger evidence (Kass & Raftery, 1995). Our results revealed a weak positive association between participants' age and their metacognitive efficiency ($rho = .13$, $p < .001$; $BF_{10} = 2780.6$), leading us to reject the hypothesis.

Supplement S5: Additional Analyses

When examining the political antecedents of metacognitive efficiency using the data from Garrett and Bond (2021), three analytical decisions need to be made, resulting in 12 potential analyses: 2 (measure of metacognitive efficiency: Mratio vs. Mdiff) x 2 (statement coding: concordance vs. congruency) x 3 (political views: 7-point ideology vs. political party vs. liberal/conservative pooled). The main finding of this study—that people on the political right exhibit metacognitive blind spots for discordant statements—is found in 11 out of these 12 possible analyses (see the results reported in detail below). The only analysis where the main pattern of results was not observed used Mdiff as a measure of metacognitive efficiency. Because the literature predominantly relies on Mratio (e.g., Lisi, 2023; Palmer et al., 2014; Said et al., 2023; Xue et al., 2021), we had preregistered the analyses using Mratio (as opposed to Mdiff). Below, we present results for all of the possible analyses—except from the one reported in the main manuscript—to provide a methodologically robust and comprehensive view of our results on political knowledge, metacognitive sensitivity, and metacognitive efficiency as a function of political party and ideology.

d', meta-d', and Mratio across different conceptualizations of political views and statement subgroups

Conservatism was associated with markedly lower object-level (d') and metacognitive (meta-d') sensitivity (Fig. S3). When coding statements as the alignment of a statement's slant with a participant's ideology, liberals exhibited higher object-level (d') and metacognitive (meta-d') sensitivity for incongruent (vs. congruent) statements. Conversely, conservatives demonstrated higher metacognitive sensitivity (meta-d') for congruent statements. For metacognitive efficiency ($Mratio$), conservatives had a harder time introspecting for incongruent (vs. congruent) statements (i.e., lower $Mratio$), thus providing support for the pattern of results reported in the main manuscript.

Participants on the political left (Democrats and liberals) showed higher object-level

(d') and metacognitive (meta-d') sensitivity than those on the political right (Republicans and conservatives; Fig. S4). When coding statements as the alignment of a statement's slant, a statement's truth, and a participant's political views, we again found a strong concordance effect: Participants from both left and right showed significantly higher object-level (d') and metacognitive (meta-d') sensitivity for concordant (vs. discordant) statements. Moreover, participants across the political and ideological spectrum showed metacognitive efficiency ($Mratio$) close to the theoretical optimum. Yet again, when inspecting different statement subgroups, we observed a striking concordance effect: Participants from the political right—but not left—showed significantly lower metacognitive efficiency ($Mratio$) for discordant (vs. concordant) statements.

When coding statements as the alignment of a statement's slant and a participant's political views, participants from the left showed higher object-level (d') and metacognitive (meta-d') sensitivity for incongruent (vs. congruent) statements, while this pattern mostly reversed for participants on the right (Fig. S5). Nonetheless, the pattern of results for metacognitive efficiency ($Mratio$) remained consistent with our main finding: Participants on the right—unlike those on the left—again had a harder time introspecting on the accuracy of their truth judgments for incongruent statements, as evidenced by markedly lower metacognitive efficiency ($Mratio$).

Mdiff as an alternative measure of metacognitive efficiency

As an alternative approach to the ratio (i.e., meta-$d'/d' = Mratio$), we can also operationalize metacognitive efficiency using the difference score (i.e., meta-d' - $d' = Mdiff$; Fleming & Lau, 2014). Below, we report the same analyses for metacognitive efficiency as reported in the main manuscript and above, but this time using the difference score (i.e., $Mdiff$).

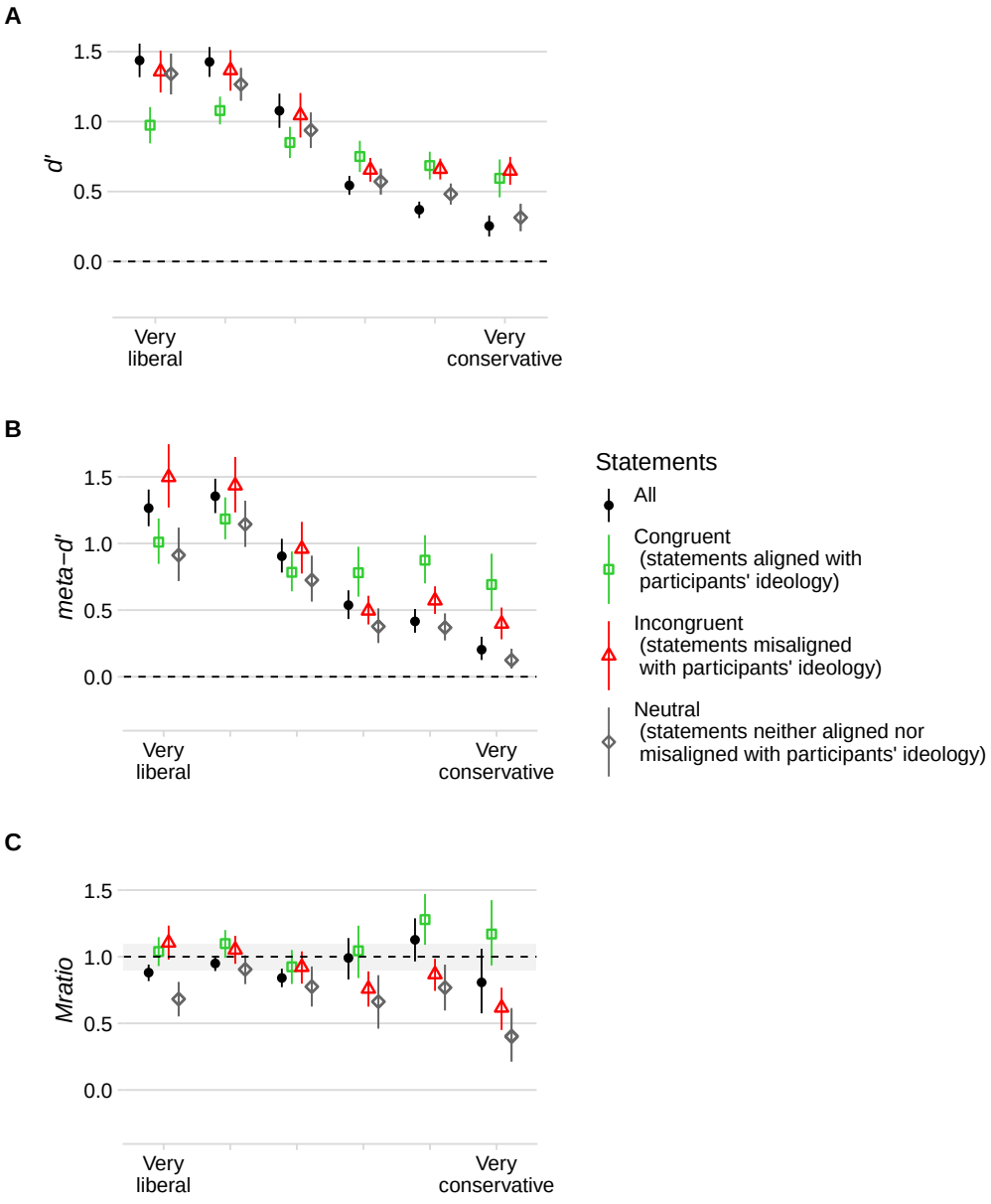
Participants across the ideological spectrum hovered around the metacognitive ideal of $Mdiff = 0$ (Fig. S6). When coding statements as the alignment of a statement's slant, a statement's truth, and a participant's ideology, we found no clear differences in

metacognitive efficiency across statement subgroups. While this finding did not confirm our main result—that conservatives showed lowered metacognitive efficiency for discordant statements—we caution against overinterpreting these $Mdiff$ results, given that $Mratio$ is the measure of metacognitive efficiency predominantly used in the literature. On the other hand, when coding statements as the alignment of a statement's slant with a participant's ideology, we again found supporting evidence for our main result: Conservatives showed significantly lower metacognitive efficiency ($Mdiff$) for incongruent statements (Fig. S7).

Overall, participants from both the political left and the political right showed metacognitive efficiency ($Mdiff$) close to the theoretical optimum, although Democrats and liberals appeared to have slightly lower metacognitive efficiency than Republicans and conservatives (Fig. S8 + S9). When coding statements as the alignment of a statement's slant, a statement's truth, and a participant's political views, we once again observed that conservatives displayed lower metacognitive efficiency for discordant statements (Fig. S8). Finally, when coding statements as the alignment of a statement's slant with a participant's political views, participants from the political right again had a harder time introspecting for incongruent statements, as reflected in markedly lower $Mdiff$ estimates (Fig. S9).

Figure S3

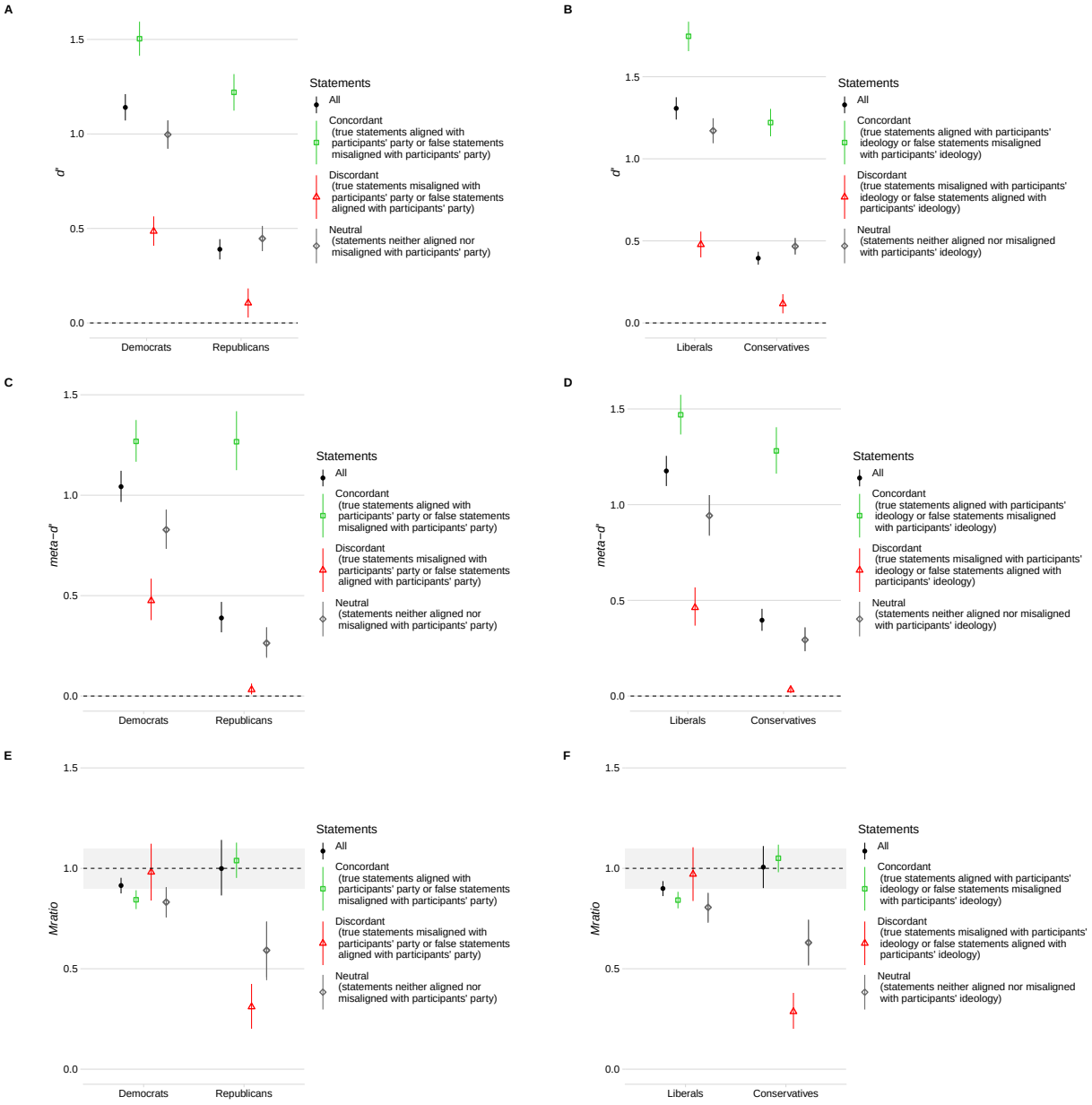
Political Knowledge (d'; *Panel A*), *Metacognitive Sensitivity (meta-d'; Panel B)*, *and Metacognitive Efficiency* (Mratio; *Panel C*) *as a Function of Political Ideology for Different Levels of Statement Congruency*



Note. Dots represent the posterior medians of the group-level averages for the parameter (d', meta-d', or $Mratio$), estimated separately for each of the six ideological subgroups and for different sets of statements. Error bars represent the 95% credible interval of the respective posterior distribution; horizontal dashed lines denote sensitivity at chance level (d' and meta-d') and theoretically optimal metacognitive efficiency ($Mratio$); the gray band for $Mratio$ marks the region of practical equivalence (ROPE; i.e., $Mratio = 0.9–1.1$).

Figure S4

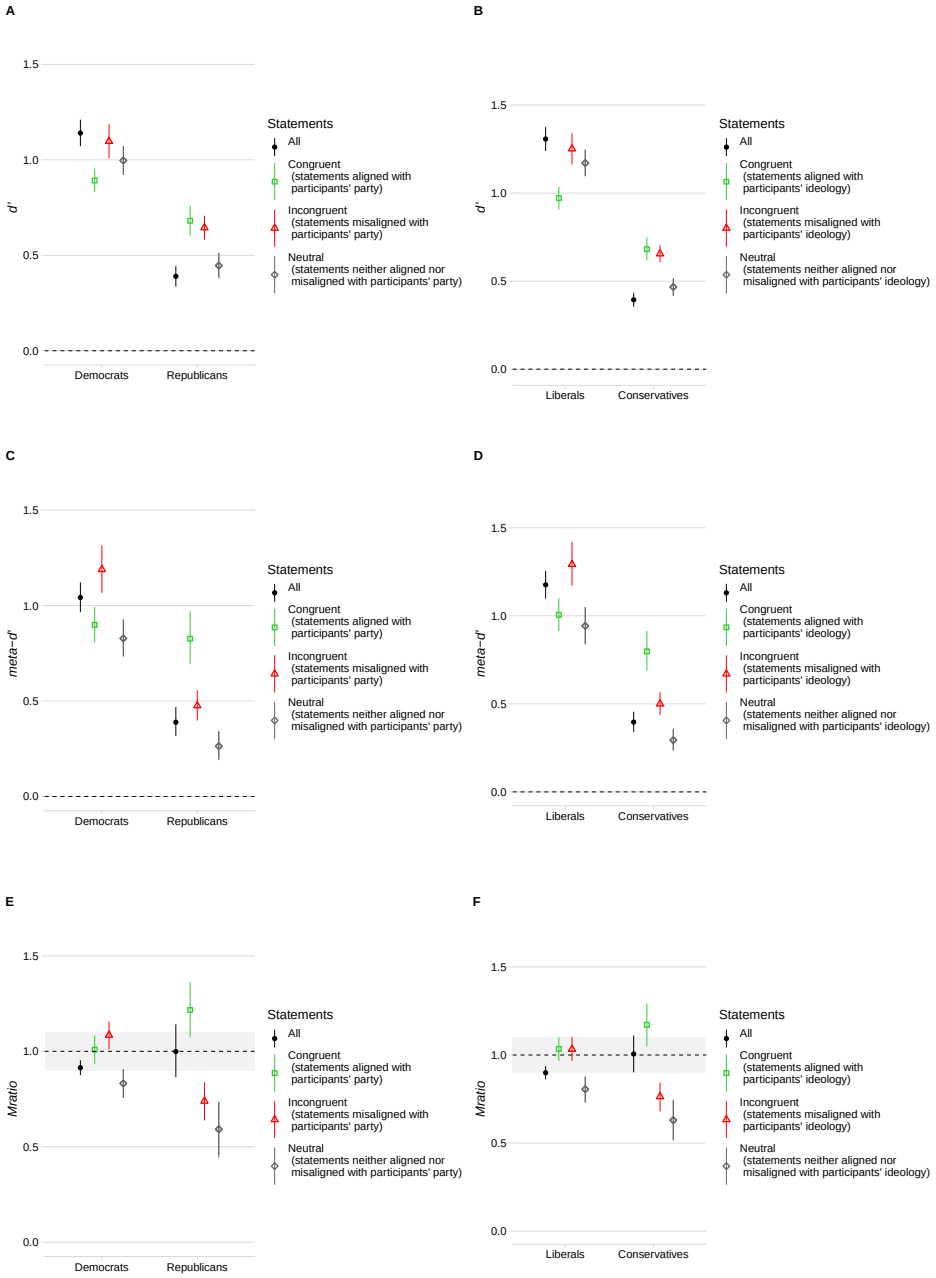
Political Knowledge (d'; *Panels A and B*), *Metacognitive Sensitivity* (*meta*-d'; *Panels C and D*), *and Metacognitive Efficiency* (Mratio; *Panels E and F*) *as a Function of Political Party and Ideology for Different Levels of Statement Concordance*



Note. Dots represent the posterior medians of the group-level averages for the parameter (d', meta-d', or $Mratio$), estimated separately for each of the party/ideological subgroups and for different sets of statements. Error bars represent the 95% credible interval of the respective posterior distribution; horizontal dashed lines denote sensitivity at chance level (d' and meta-d') and theoretically optimal metacognitive efficiency ($Mratio$); the gray band for $Mratio$ marks the region of practical equivalence (ROPE; i.e., $Mratio$ = 0.9–1.1).

Figure S5

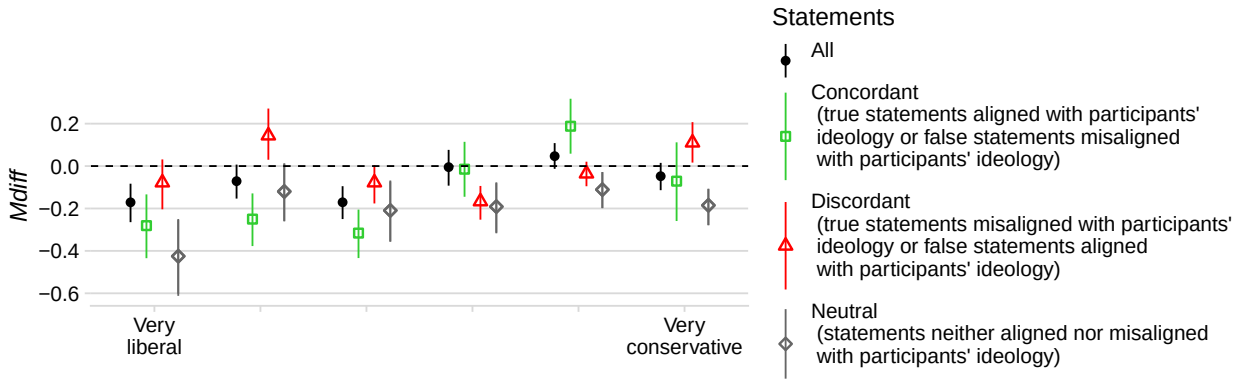
Political Knowledge (d'; *Panels A and B*), *Metacognitive Sensitivity* (meta-d'; *Panels C and D*), *and Metacognitive Efficiency* (Mratio; *Panels E and F*) *as a Function of Political Party and Ideology for Different Levels of Statement Congruency*



Note. Dots represent the posterior medians of the group-level averages for the parameter (d', meta-d', or $Mratio$), estimated separately for each of the party/ideological subgroups and for different sets of statements. Error bars represent the 95% credible interval of the respective posterior distribution; horizontal dashed lines denote sensitivity at chance level (d' and meta-d') and theoretically optimal metacognitive efficiency ($Mratio$); the gray band for $Mratio$ marks the region of practical equivalence (ROPE; i.e., $Mratio = 0.9–1.1$).

Figure S6

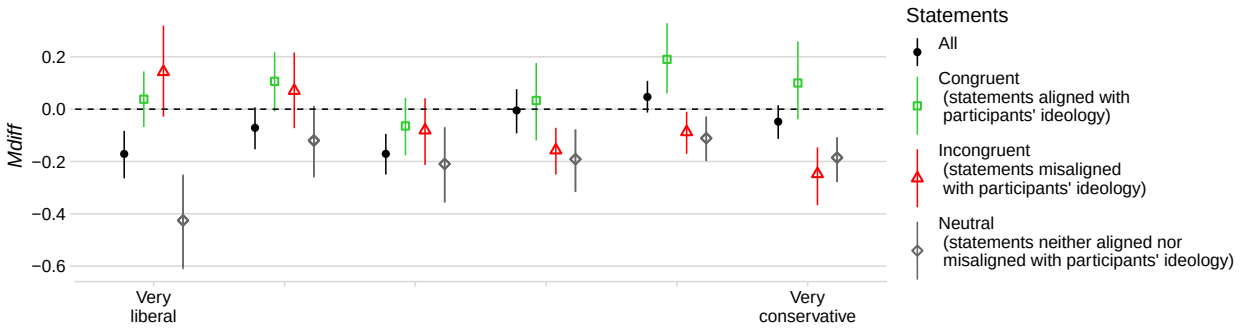
*Metacognitive Efficiency (*Mdiff*) as a Function of Political Ideology for Different Levels of Statement Concordance*



Note. Dots represent the posterior medians of the group-level averages for the parameter ($Mdiff$), estimated separately for each of the six ideological subgroups and for different sets of statements. Error bars represent the 95% credible interval of the respective posterior distribution; the horizontal dashed line denotes theoretically optimal metacognitive efficiency.

Figure S7

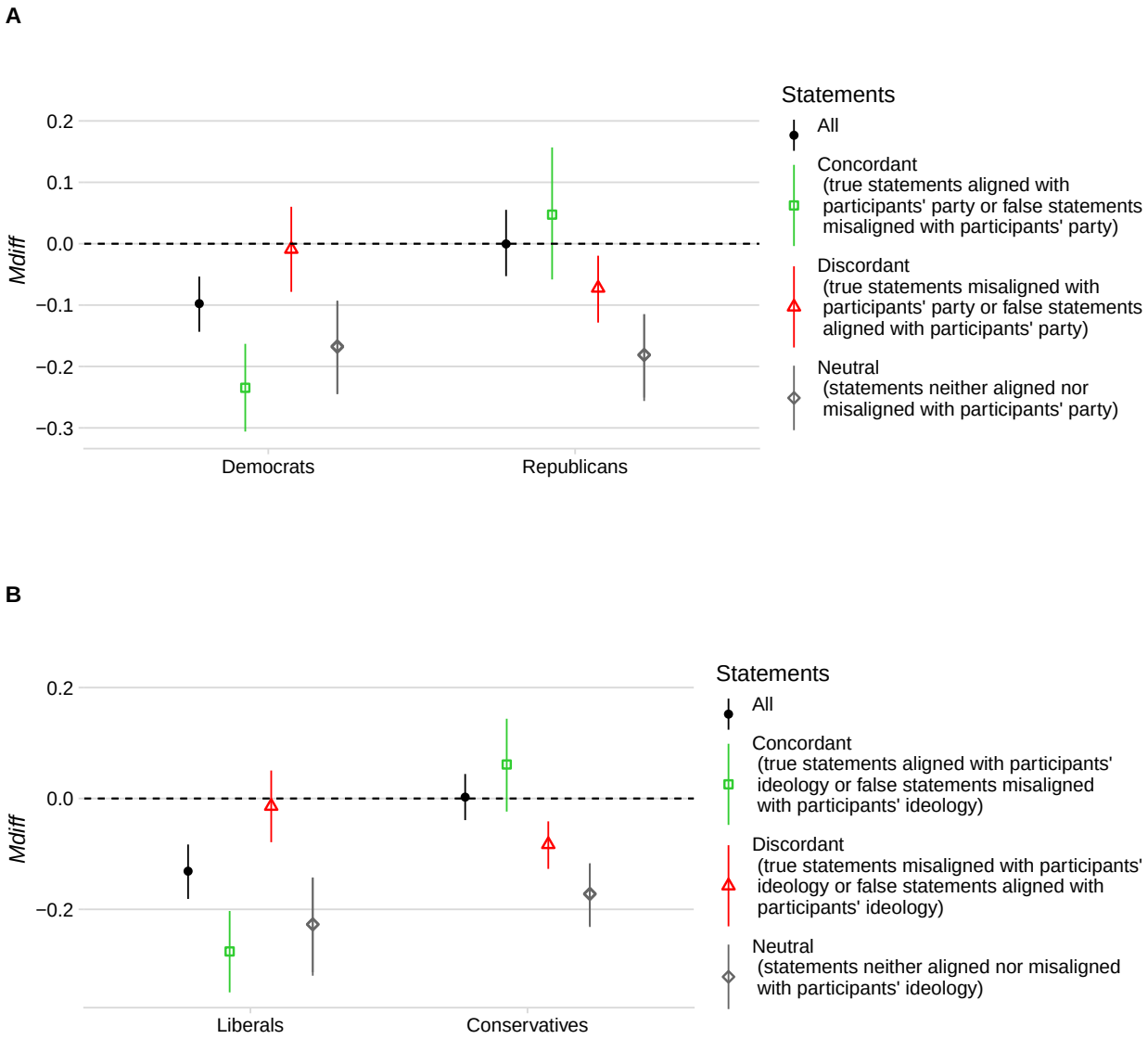
*Metacognitive Efficiency (*Mdiff*) as a Function of Political Ideology for Different Levels of Statement Congruency*



Note. Dots represent the posterior medians of the group-level averages for the parameter ($Mdiff$), estimated separately for each of the six ideological subgroups and for different sets of statements. Error bars represent the 95% credible interval of the respective posterior distribution; the horizontal dashed line denotes theoretically optimal metacognitive efficiency.

Figure S8

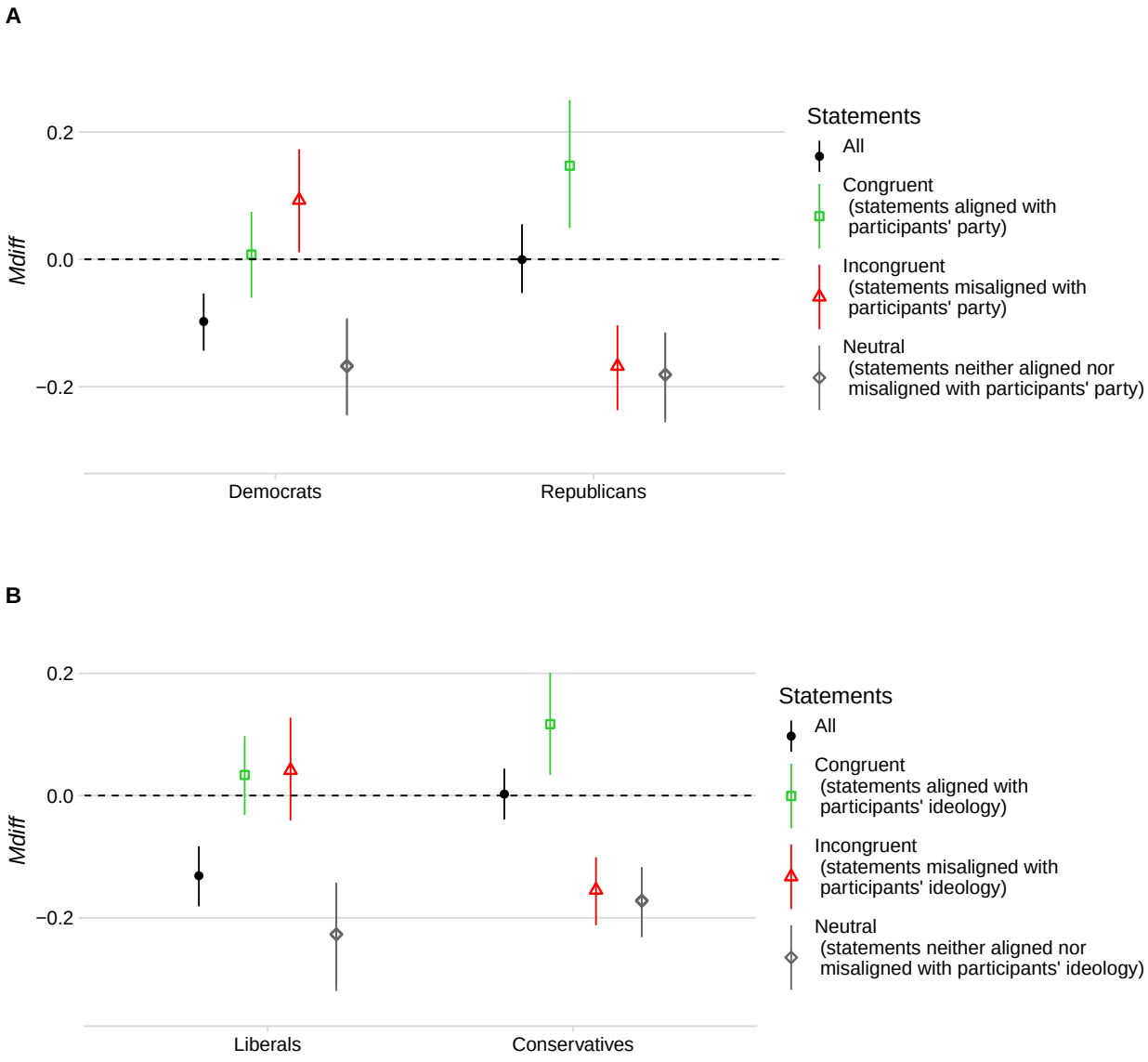
*Metacognitive Efficiency (*Mdiff*) as a Function of Political Party (Panel A) and Ideology (Panel B) for Different Levels of Statement Concordance*



Note. Dots represent the posterior medians of the group-level averages for the parameter ($Mdiff$), estimated separately for each of the party/ideological subgroups and for different sets of statements. Error bars represent the 95% credible interval of the respective posterior distribution; horizontal dashed lines denote theoretically optimal metacognitive efficiency.

Figure S9

*Metacognitive Efficiency (*Mdiff*) as a Function of Political Party (Panel A) and Ideology (Panel B) for Different Levels of Statement Congruency*



Note. Dots represent the posterior medians of the group-level averages for the parameter ($Mdiff$), estimated separately for each of the party/ideological subgroups and for different sets of statements. Error bars represent the 95% credible interval of the respective posterior distribution; horizontal dashed lines denote theoretically optimal metacognitive efficiency.

Supplement S6: Robustness Checks

Group-Level

We investigated group-level estimates of d', meta-d', and $Mratio$ across various subgroups of statements (Fig. S10). Overall, participants demonstrated moderate-low ability to distinguish between true and false news (Fig. S10, Panel A). They discerned recognized (vs. unrecognized) and non-satire (vs. satire) statements at a higher rate. We also investigated d' results based on statement concordance. When conceptualized as a statement's slant and truth aligning with a participant's party/ideology, participants did better at detecting concordant (vs. discordant) statements. By contrast, when conceptualized as a statement's slant only aligning with a participant's party/ideology, there were no noticeable differences between concordant and discordant statements. A very similar pattern of results emerged for participants' meta-d' (Fig. S10, Panel B). Lastly, participants' metacognitive efficiency for all statements hovered around the theoretically optimal value (i.e., $Mratio = 1$; Fig. S10, Panel C). $Mratio$ was slightly higher for recognized (vs. unrecognized) statements, and substantially higher for satire (vs. non-satire) statements. We also investigated $Mratio$ results based on statement concordance. When conceptualized as a statement's slant and truth aligning with a participant's party/ideology, participants showed higher $Mratio$ for concordant (vs. discordant) statements. Similarly, when conceptualized as a statement's slant only aligning with a participant's party/ideology, $Mratio$ was higher for concordant (vs. discordant) statements.

Individual-Level

Participant Attrition

We investigated individual-level estimates of $Mdiff$ (used to test H3.1–H3.4) across different subgroups of participant completion rates; Fig. S11). Overall, no substantial differences emerged among all participants, participants who completed six or more waves, and participants who completed 12 waves. Yet, it is worth noting that $Mdiff$ appeared to

increase slightly with wave completion rates by tendency.

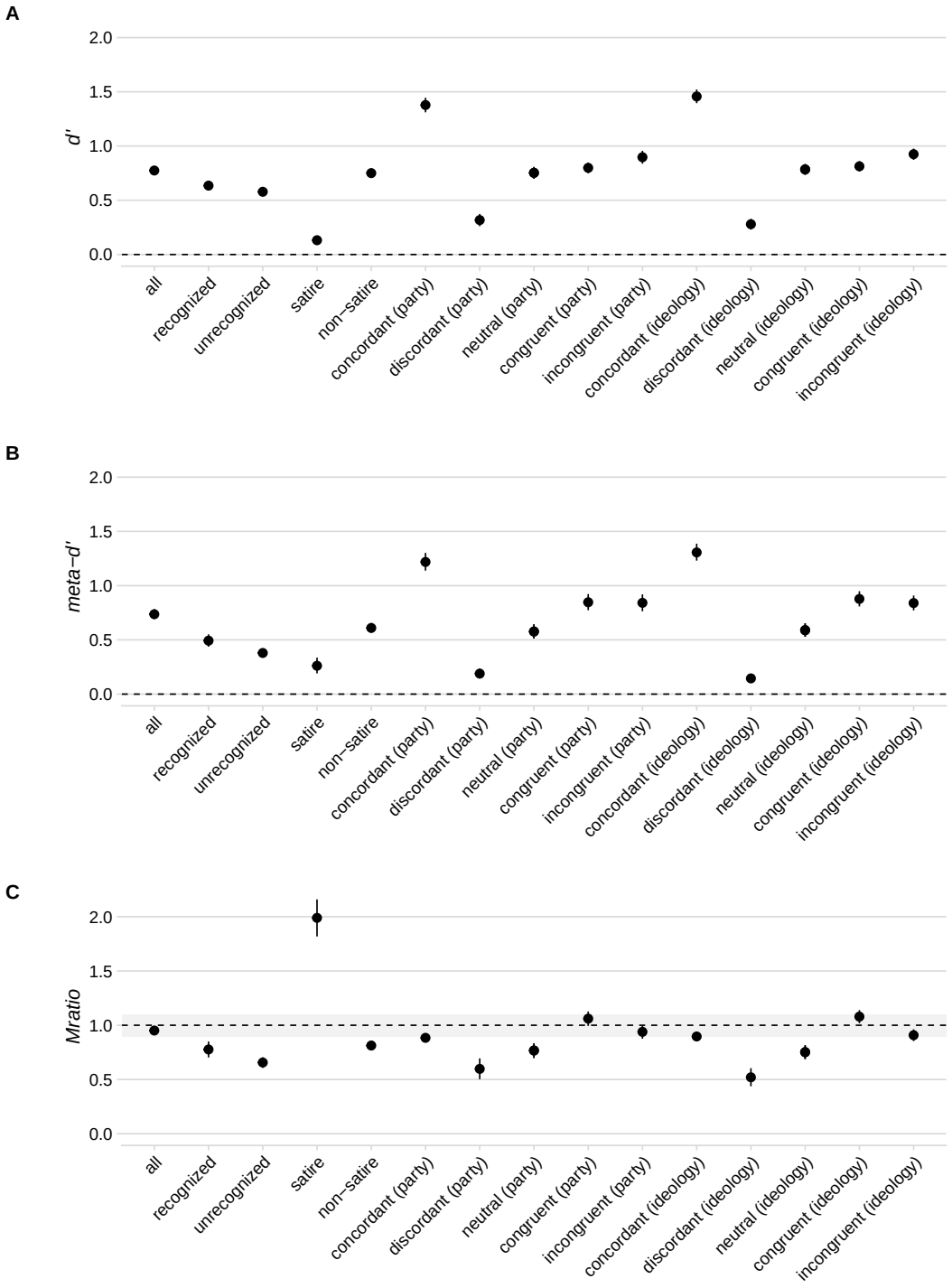
Statement Characteristics

We investigated individual-level estimates of $Mdiff$ (used to test H3.1–H3.4) across different subgroups of statements (Fig. S12). Specifically, we compared different estimates of $Mdiff$ within participants based on concordance, recognition, and satire. Here, we conceptualized statement concordance as a combination of (i) a statement's truth and (ii) the alignment of its political slant with a participant's political party. We found the following associations between participants' individual-level estimates of $Mdiff$ across different statement subgroups:

- Participants' individual-level estimates of $Mdiff$ for concordant and discordant statements were not significantly correlated ($rho = -.01$, $p = .795$; $BF_{10} = .10$).
- Participants' individual-level estimates of $Mdiff$ for concordant and neutral statements were significantly positively correlated ($rho = .23$, $p < .001$; $BF_{10} = 3.98 \times 10^7$).
- Participants' individual-level estimates of $Mdiff$ for discordant and neutral statements were significantly positively correlated ($rho = .09$, $p = .011$; $BF_{10} = 2.23$).
- Participants' individual-level estimates of $Mdiff$ for recognized and unrecognized statements were significantly positively correlated ($rho = .11$, $p < .001$; $BF_{10} = 41.9$).
- Participants' individual-level estimates of $Mdiff$ for statements originating in satire and non-satire were significantly positively correlated ($rho = .11$, $p < .001$; $BF_{10} = 48.5$).

Figure S10

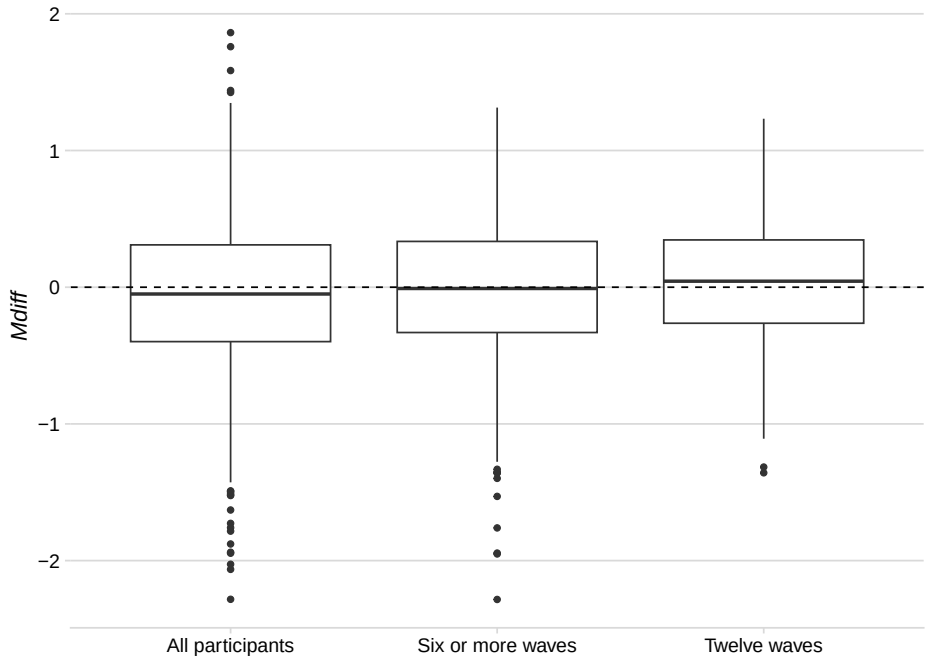
*Political Knowledge (*d'*; Panel A), Metacognitive Sensitivity (meta-*d'*; Panel B), and Metacognitive Efficiency (*Mratio*; Panel C) across Different Subgroups of Statements*



Note. Dots represent the posterior medians of the group-level averages for the parameter (d', meta-d', or $Mratio$), estimated separately for each of the statement subgroups. Error bars represent the 95% credible interval of the respective posterior distribution; horizontal dashed lines denote sensitivity at chance level (d' and meta-d') and theoretically optimal metacognitive efficiency ($Mratio$). The gray band for $Mratio$ marks the region of practical equivalence (ROPE; i.e., $Mratio$ = 0.9–1.1).

Figure S11

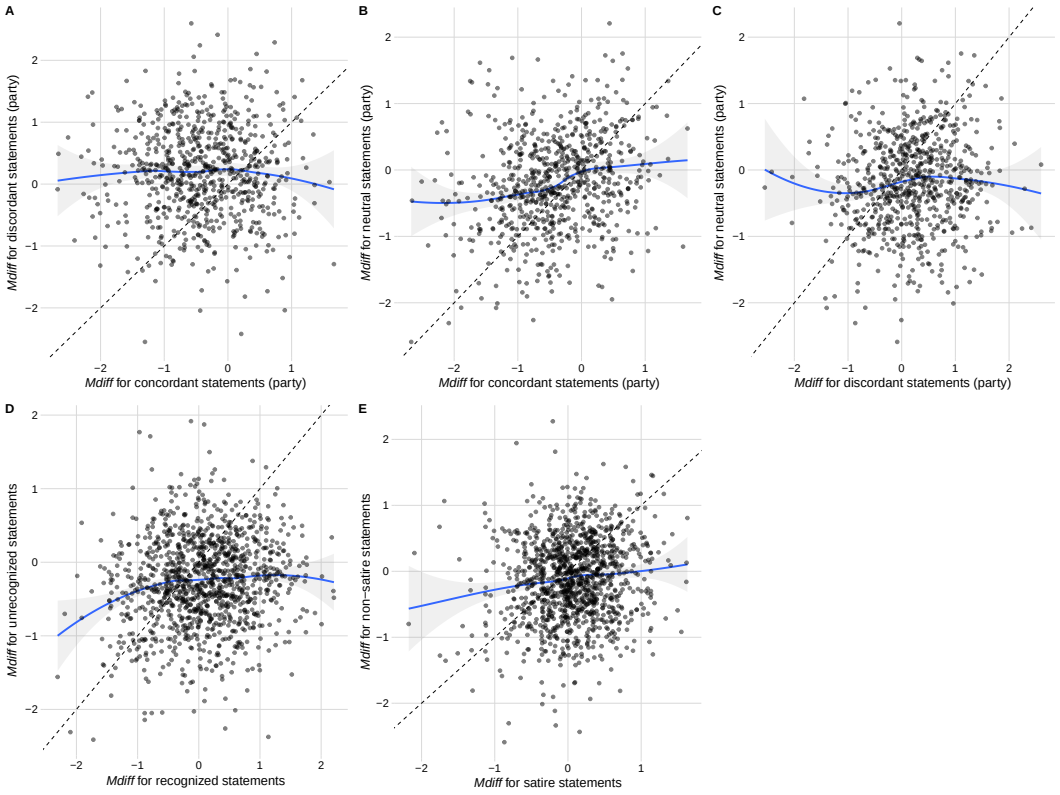
Mdiff *Estimated at the Individual Level across Different Subgroups of Participant Completion Rates*



Note. The horizontal dashed line at $Mdiff = 0$ denotes theoretically optimal metacognitive efficiency.

Figure S12

Mdiff *Estimated at the Individual Level across Different Subgroups of Statements*



Note. Curves are robust LOESS smooths with 95% confidence intervals (shaded). The dashed identity lines denote perfect correlation for $Mdiff$ across different statement subgroups.

References

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science*, *17*(1), 78–98. https://doi.org/10.1177/1745691620986135

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *NeuroImage*, *73*, 80–94. https://doi.org/https://doi.org/10.1016/j.neuroimage.2013.01.054

Derreumaux, Y., Shamsian, K., & Hughes, B. L. (2023). Computational underpinnings of partisan information processing biases and associations with depth of cognitive reasoning. *Cognition*, *230*, Article 105304. https://doi.org/10.1016/j.cognition.2022.105304

Fischer, H., Amelung, D., & Said, N. (2019). The accuracy of German citizens' confidence in their climate change knowledge. *Nature Climate Change*, *9*(10), 776–780. https://doi.org/10.1038/s41558-019-0563-0

Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1), Article nix007. https://doi.org/10.1093/nc/nix007

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, Article 443. https://doi.org/10.3389/fnhum.2014.00443

Garrett, R. K., & Bond, R. M. (2021). Conservatives' susceptibility to political misperceptions. *Science Advances*, *7*(23), Article eabf1234. https://doi.org/10.1126/sciadv.abf1234

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Lisi, M. (2023). Navigating the COVID-19 infodemic: The influence of metacognitive efficiency on health behaviours and policy attitudes. *Royal Society Open Science*, *10*(9), Article 230417. https://doi.org/https://doi.org/10.1098/rsos.230417

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press. https://doi.org/10.4324/9781410611147

Palmer, E. C., David, A. S., & Fleming, S. M. (2014). Effects of age on metacognitive efficiency. *Consciousness and Cognition*, *28*, 151–160. https://doi.org/10.1016/j.concog.2014.06.007

R Core Team. (2023). *R: A language and environment for statistical computing*. Foundation for Statistical Computing. Vienna, Austria.

Rabbitt, P., & Vyas, S. (1981). Processing a display even after you make a response to it. How perceptual errors can be corrected. *The Quarterly Journal of Experimental Psychology*, *33*(3), 223–239. https://doi.org/10.1080/14640748108400790

Rausch, M., & Zehetleitner, M. (2016). Visibility is not equivalent to confidence in a low contrast orientation discrimination task. *Frontiers in Psychology*, *7*, Article 591. https://doi.org/10.3389/fpsyg.2016.00591

Said, N., Potinteu, A. E., Brich, I., Buder, J., Schumm, H., & Huff, M. (2023). An artificial intelligence perspective: How knowledge and confidence shape risk and benefit perception. *Computers in Human Behavior*, 107855. https://doi.org/10.1016/j.chb.2023.107855

Scott, R. B., Dienes, Z., Barrett, A. B., Bor, D., & Seth, A. K. (2014). Blind insight: Metacognitive discrimination despite chance task performance. *Psychological Science*, *25*(12), 2199–2208. https://doi.org/10.1177/0956797614553944

Srivastava, S. (2018). *Sound inference in complicated research: A multi-strategy approach.* PsyArXiv. https://doi.org/10.31234/osf.io/bwr48

van den Akker, O. R., Weston, S., Campbell, L., Chopik, B., Damian, R., Davis-Kean, P., Hall, A., Kosie, J., Kruse, E., Olsen, J., Ritchie, S., Valentine, K. D., van't Veer, A., & Bakker, M. (2021). Preregistration of secondary data analysis: A template and tutorial. *Meta-Psychology*, *5*, Article MP.2020.2625. https://doi.org/10.15626/MP.2020.2625

Xue, K., Shekhar, M., & Rahnev, D. (2021). Examining the robustness of the relationship between metacognitive efficiency and metacognitive bias. *Consciousness and Cognition*, *95*, 103196. https://doi.org/10.1016/j.concog.2021.103196