Spotting False News and Doubting True News: A Meta-Analysis of News Judgements

Jan Pfänder[1] & Sacha Altay[2]

[1] Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, PSL University, CNRS, France

[2] Department of Political Science, University of Zurich, Switzerland

Abstract

How good are people at judging the veracity of news? We conducted a systematic literature review and pre-registered meta-analysis of 173 effect sizes from 48 experimental papers evaluating accuracy ratings of true and false news ($N_{participants} = 100777$ from 30 countries across 6 continents). We found that people rated true news as much more accurate than false news (d = 1.01 [0.9, 1.12]) and were slightly better at rating false news as false than at rating true news as true (d = 0.28 [0.2, 0.35]). In other words, participants were able to discern true from false news, and were slightly more skeptical than gullible. The political concordance of the news had no effect on discernment, but participants were more skeptical of politically discordant news. These findings lend support to crowdsourced fact-checking initiatives, and suggest that, to improve discernment, there is more room to increase the acceptance of true news than to reduce the acceptance of false news.

Keywords: Misinformation; fake news; false news; news judgment; news accuracy; news discernment

JP received funding from the SCALUP ANR grant ANR-21-CE28-0016-01

SA received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement nr. 883121).

The authors made the following contributions. Jan Pfänder: Conceptualization, Data collection, Statistical Analyses, Writing - Original Draft, Writing - Review & Editing; Sacha Altay: Conceptualization, Data collection, Writing - Original Draft, Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to Sacha Altay. E-mail: sacha.altay@gmail.com

Introduction

Many have expressed concerns that we live in a “post-truth” era, and that people cannot tell the truth from falsehoods anymore. In parallel, populist leaders around the world have tried to erode trust in the news by delegitimizing journalists and media outlets. Since the 2016 US presidential election, over 4000 scientific articles have been published on the topic of false news. Across the world, numerous experiments evaluating the effect of interventions against misinformation or susceptibility to misinformation have relied on a similar design feature: having participants rate the accuracy of true and false headlines–typically in a Facebook-like format, with an image, title, lede, and source, or as an isolated title/claim. Taken together, these studies allow us to shed some light on the most common fears voiced about false news, namely that people may fall for false news, distrust true news, or may be unable to discern between true and false news. In particular, we investigated whether people rate true news as more accurate than false news (discernment) and whether they were better at rating false news as inaccurate than at rating true news as accurate (response bias). We also investigated various moderators of discernment and response bias such as political congruence, the topic of the news, or the presence of a source.

Establishing whether people can spot false news is important to design interventions against misinformation: if people lack the skills to spot false news, interventions should be targeted at improving skills to detect false news, whereas if people have the ability to spot false news but nonetheless engage with it, the problem lies elsewhere and may be one of motivation or (in)attention that educational interventions may struggle to address.

Past work has reliably shown that people do not fare better than chance at detecting lies because most verbal and non-verbal cues people use to detect lies are unreliable[1]. Why would this be any different for detecting false news? People make snap judgments to evaluate the quality of the news they come across[2], and rely on seemingly imperfect proxies such as the source of information, police and fonts, the presence of hyperlinks, the quality of visuals, ads, or the tone of the text[3,4]. In experimental settings, participants report relying on intuitions and tacit knowledge to judge the accuracy of news headlines[5]. Yet, a scoping review of the literature on belief in false news (including a total of 26 articles) has shown that, in experiments, participants “can detect deceitful messages reasonably well”[6]. Similarly, a survey on 150 misinformation experts has shown that 53% of experts agree that “people can tell the truth from falsehoods” – while only 25% of experts disagreed with the statement[5]. Unlike the unreliable proxies people rely on to detect lies in interpersonal contexts, there are reasons to believe that some of the cues people use to detect false news may, on average, be reliable. For instance, the news outlets people trust the less do publish lower quality news and more false news, as people’s trust ratings of news outlets correlate strongly with fact-checkers ratings in the US and Europe[7,8]. Moreover, false news share some distinctive properties, such as being more politically slanted[9], being more novel, surprising, or disgusting, being more sensationalist, funnier, less boring, and less negative[10,11], or being more interesting-if-true[12]. These features aim at increasing engagement, but they do so at the expense of accuracy, and in many cases people may pick up on it. This led us to pre-register the hypothesis that people would rate true news as more accurate than false news. Yet, legitimate concerns have been raised about the lack of data outside of the US, especially in some Global South countries where the misinformation problem is arguably worst. Our meta-analysis covers 30 countries across 6 continents and directly addresses concerns about the over-representation of US-data.

H1: People rate true news as more accurate than false news.

While many fear that people are exposed to too much misinformation, too easily fall for it, and are overly influenced by it, a growing body of researchers is worried that people are exposed to too little reliable information, commonly reject it, and are excessively resistant to it[13,14]. Establishing whether true news skepticism (excessively rejecting true news) is of similar magnitude to false news gullibility (excessively accepting false news) is important for future studies on misinformation: if people are excessively gullible, interventions should primarily aim at fostering skepticism, whereas if people are excessively skeptical, interventions should focus on increasing trust in reliable information. For these reasons, in addition to investigating discernment (H1), we also looked at response bias by comparing the magnitude of true news skepticism to false news gullibility. Research in psychology has shown that people exhibit a "truth bias"[15,16], such that they tend to accept incoming statements rather than rejecting them. Similarly, work on interpersonal communication has shown that, by default, people tend to accept communicated information[17]. However, there are reasons to be skeptical that the truth-default-theory holds for news judgments. It has been hypothesized that people display a truth bias in interpersonal contexts because information in these contexts is, in fact, often true[15]. When it comes to the news judgments, it is not clear that people by default expect news stories to be true. Trust in the news and journalists is low worldwide[18], and a significant part of the population hold cynic views of the news[19]. Similarly, populists leaders across the world have attacked the credibility of the news media and instrumentalized the concept of fake news to discredit quality journalism[20,21]. Disinformation strategies such as "flooding the zone" with false information[22,23] have been shown to increase skepticism in news judgments[5]. Moreover, in many studies included in our meta-analysis, the news stories were presented in a social media format (most often Facebook), which could fuel skepticism in news judgments. Indeed, people trust news[2] and information more generally[24] less on social media than on news websites. In line with these observations, some empirical evidence suggests that for news judgments, people display the opposite of a truth bias[25], namely a conservative response bias, whereby people tend to rate all news as more false than they are[5,26,27]. We thus predicted that when judging the accuracy of news, participants will err on the side of skepticism more than on the side of gullibility. Precisely, we predicted that people will be better at rating false news as false than rating true news as true.

H2: People are better at rating false news as false than true news as true.

Finally, we investigated potential moderators of H1 and H2, such as the country where the experiment was conducted, the format of the news headlines, the topic, whether the source of the news was displayed, and political concordance of the news. Past work has suggested that displaying the source of the news has a small effect at best on accuracy ratings[28], whereas little work has investigated differences in news judgements across countries, topics, and format. The effect of political concordance on news judgements is debated. Participants may be motivated to believe politically congruent (true and false) news, motivated to disbelieve politically incongruent news, or not be politically motivated at all but still display such biases[29]. We formulated research questions instead of hypotheses for our moderator analyses because of a lack of strong theoretical expectations.

The present study

We conducted a systematic literature review and pre-registered meta-analysis based on 48 publications, providing data on 131 samples (100777 participants) and 173 effects (i.e. **k**, the meta analytic observations)[1]. For a publication to be included in our meta-analysis, we set seven eligibility criteria: (1) We considered as relevant all document types with original data (not only published ones, but also reports, pre-prints and working papers). When different publications were using the same data, a scenario we encountered several times, we included only one publication (which we picked arbitrarily). (2) We only included articles that measured perceived accuracy (including “accuracy”, “credibility”, “trustworthiness”, “reliability” or “manipulativeness”), and (3) did so for both true and false news. (4) We only included studies relying on real-world news items. Accordingly, we excluded studies in which researchers made up the false news items, or manipulated the properties of the true news items. (5) We could only review articles that we could access (although it was almost never an issue) and (6) articles that provided us with the relevant summary statistics (means and standard deviations for both false and true news), or publicly available data that allowed us to calculate those. In cases where we were not able to retrieve the relevant summary statistics either way, we contacted the authors. (7) Finally, to ensure comparability, we only included studies that provided a neutral control condition[2]. After starting the literature search, we added further search criteria in order to diminish the vast number of results (see methods). Rejection decisions for all retrieved papers are documented and can be accessed on the OSF project page.

We found that on average people are good at discerning true from false news, and rate true news as much more accurate than false news. However, they are slightly better at rating false news as inaccurate than at rating true news as accurate.

Results

Descriptives

Our meta-analysis includes publications from 30 countries and 6 continents. However, the number of participants varies a lot, led by the United States with 46166 participants

[1] Sometimes a sample provided several effect sizes, for example, when separate accuracy ratings are available by news topic, or when follow-up studies were conducted on the same participants. We account for the resulting hierarchical structure of the data in our statistical models.

[2] For example,[30], among other things, test the effect of an interest prime vs. an accuracy prime. A neutral control condition - one that is comparable to those of other studies - would have been no prime at all. We therefore excluded the paper.

(46% of all participants; see Fig. 1 for number of effect sizes per country). The average sample size was 769.29 (min = 53, max = 9474, median = 412)[3].

In total, participants rated the accuracy of 1743 unique news items. On average, a participant rated 17.24 news items per study (min = 6, max = 62, median = 18). For 46 samples, news items were sampled from a pool of news (the pool size ranged from 12 to 255, with an average pool size of 58.82 items). The vast majority of studies (168 out of 173 effects) used a within participant design for manipulating news veracity, with each participant rating both true and false news items. Almost all effect sizes are from online studies (165 out of 173).

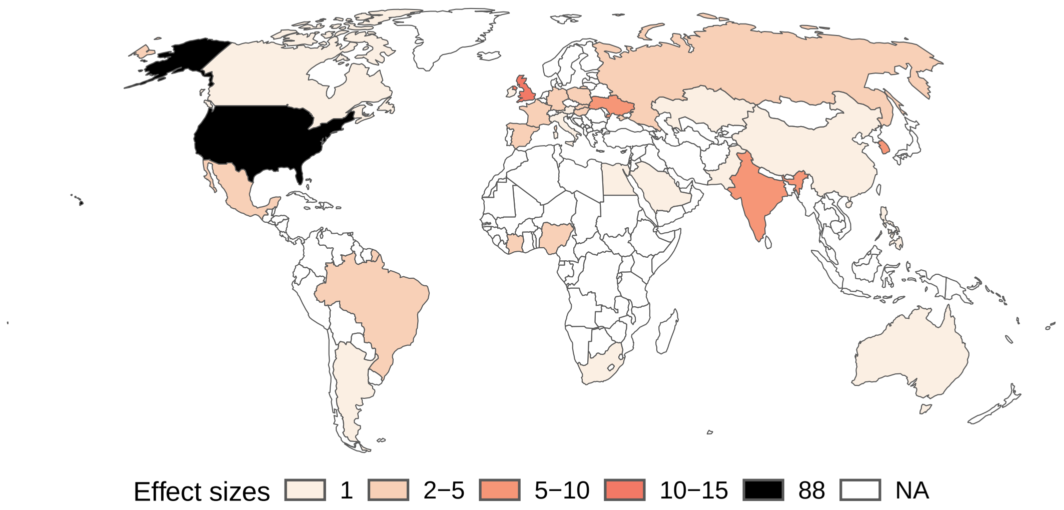

Figure 1. A map of the number of effect sizes per country.

Analytic procedures. All analyses were pre-registered and the choice of models was informed by simulations we conducted before having the data. To test H1, we calculated a discernment score by subtracting the mean accuracy ratings of false news from the mean accuracy ratings of true news, such that higher scores indicate better discernment. To test H2, we first calculated a judgment error for true and fake news respectively. Generally, we define error as the distance of observed judgment performance of a sample to the best possible performance (see Fig. 2). For false news, error is defined as the distance between the accuracy score and the bottom of the scale. For instance, for an average false news

[3] In rare cases we did not find information on exactly how many participants were in each experimental condition. In such cases, we took the overall reported sample size and assumed an even split across conditions (i.e. dividing the overall N by the number of conditions).

accuracy rating of `2.2` on a 4-point accuracy scale going from `1`, not accurate at all, to `4`, completely accurate, the error would be: `2.2 - 1 = 1.2`. For true news, error is defined as the distance between the accuracy value and the top of the scale. For instance, for an average true news accuracy rating of `2.5` on a 4-point accuracy scale, the error would be: `4 - 2.5 = 1.5`. We then calculate the response bias as the difference between the two errors, subtracting the true news error score from the false news error score (e.g. response bias: 1.5 - 1.2 = 0.3).

Fig. 2 offers a descriptive (irrespective of sample sizes) overview of all accuracy ratings that we calculated our effect sizes from. On average, true news were rated as more accurate than false news, as shown by the positive discernment score (0.28). We also see that false news discrimination is better than true news discrimination, i.e., the distance between true news ratings and the top of the scale (0.40) is greater than the distance between false news ratings and bottom of the scale (0.32), yielding a positive response bias (0.40 - 0.32 = 0.08).

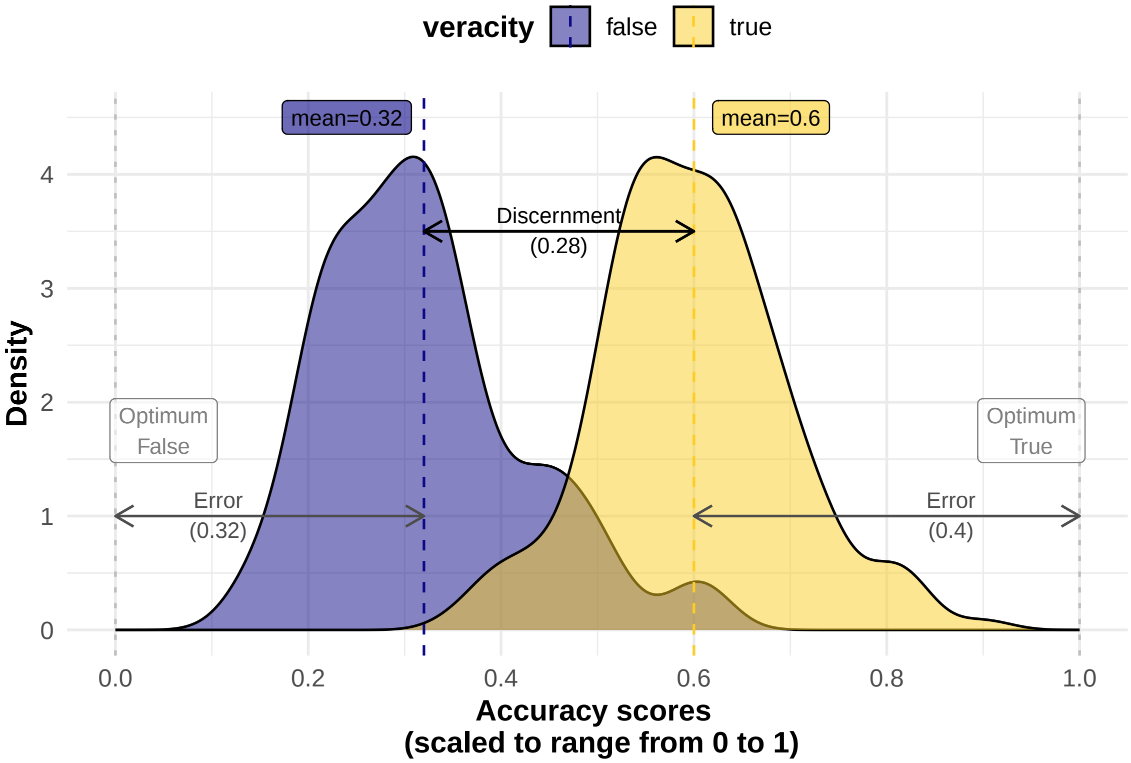

Figure 2. Distribution of accuracy ratings for true and false news, scaled to range from 0 to 1. The figure illustrates discernment (the distance between the mean for true news and the mean for false news) and the errors (distance to right end for true news, to left end for false news) from which the response bias is computed. A larger error for true news compared to false news yields a positive response bias. In this descriptive figure, unlike in the meta-analysis, ratings and effect sizes are not weighted by sample size.

Discrimination and response bias can only be (meaningfully) computed on scales using symmetrical labels (e.g., "True" vs "False" or "Definitely fake" [1] to "Definitely real" [7]).

55% of effects included in the meta-analysis used scales with perfectly symmetrical labels, while 37% used imperfectly symmetrical scale labels (e.g., [1] not at all accurate, [2] not very accurate, [3] somewhat accurate, [4] very accurate)[4]. We do not find a difference regarding response bias between effects from studies using perfectly and imperfectly symmetrical scales, but we do find a difference regarding discernment (see Appendix B). While studies with perfectly symmetric scales tend to yield lower discernment scores, our results hold across both types of scale.

We calculated standardized mean changes using change score standardization (SMCC) to compare effect sizes across studies[31]. This measure expresses effects in units of (pooled) standard deviations, allowing for comparison across different scales while accounting for the statistical dependence between true and false news ratings arising from the within participant design used by most studies (168 out of 173 effect sizes). For example, an SMCC of one for discernment would mean that people rate true news as more accurate by one pooled standard deviation, with that standard deviation taking into account the correlation between true and false news ratings (see methods).

For the remaining 5 effect sizes from studies that used a between participant design, we calculated Hedge's g, a common measure of standardized mean difference that assumes independence between groups[32]. In Appendix A, we show that our results hold across alternative effect measures, the SMCC yielding the most conservative estimates. There, we also provide effect estimates in units of the original scales separately for each scale.

We used multilevel meta models with clustered standard errors at the sample level to account for cases in which the same sample contributed various effect sizes (i.e. the meta-analytic units of observation).

Main results.

Discernment (H1). Supporting H1, participants rated true news as more accurate than false news on average. Pooled across all studies, the average discernment estimate is large (d = 1.01 [0.9, 1.12]). As shown in Fig. 3, 169 of 173 estimates are positive. Of the positive estimates, 3 have a confidence interval that includes 0, as do 2 of the negative estimates. Most of the variance in the effect sizes observed above is explained by between-sample heterogeneity ($I2_{between} = 92.57\%$). Within-sample heterogeneity is comparatively small ($I2_{within} = 7.4\%$), indicating that when the same participants were observed on several occasions (i.e. the same sample contributed several effect sizes), on average, discernment performance was similar across those observations. The share of the variance attributed to sampling error is very small (0.04%), which is indicative of the large sample sizes and thus precise estimates.

Response bias (H2). We found support for H2, with participants being better at rating false news as inaccurate than at rating true news as accurate (i.e. false news discrimination was on average higher than true news discrimination). However, the average response bias estimate is relatively small (d = 0.28 [0.2, 0.35]).

[4] We could only compute this variable for scales that explicitly labeled each scale point, resulting in missing values for 8% of effects.

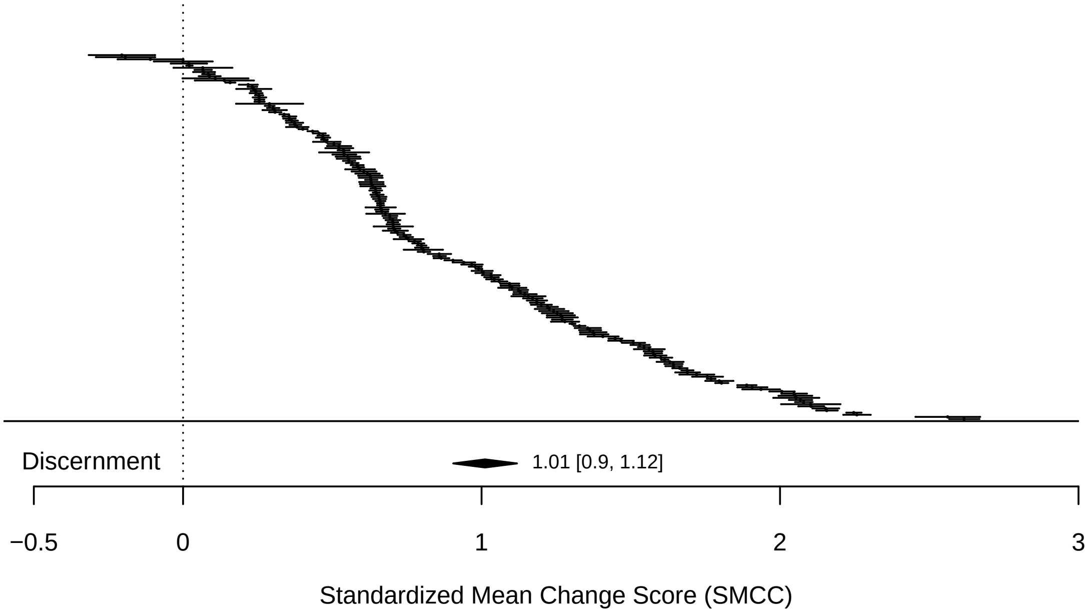

Figure 3. Forest plot for discernment. Effects are weighed by their sample size. Horizontal bars represent 95% confidence intervals. The average estimate is the result of a multilevel meta model with clustered standard errors at the sample level.

As shown in Fig 4, 137 of 173 estimates are positive. Of the positive estimates, 4 have a confidence interval that includes 0, as do 5 of the negative estimates.

By contrast with discernment, the largest share of variance of the effects for response bias is explained by within-sample heterogeneity ($I2_{within} = 63.88\%$; $I2_{between} = 36.06\%$; sampling error = 0.05%). Whenever we observe within sample variation in our data, it is because several effects were available for the same sample. This is mostly the case for studies with multiple survey waves, or when effects were split by different news topics, suggesting that these factors may account for some of that variation. In our moderator analyses below, we compare across samples and broad categories, thereby glossing over much of that within variation. An exception is political concordance, a factor that has generally been manipulated within samples and in similar ways across studies.

Moderators. Following the pre-registered analysis plan, we ran a separate meta regression for each moderator by adding the respective moderator variable as a fixed effect to the multilevel meta models. We report regression tables and visualizations in Appendix B.

Cross-cultural variability. For samples based in the United States (88/173 effect sizes), discernment was higher than for samples based in other countries, on average (Δ

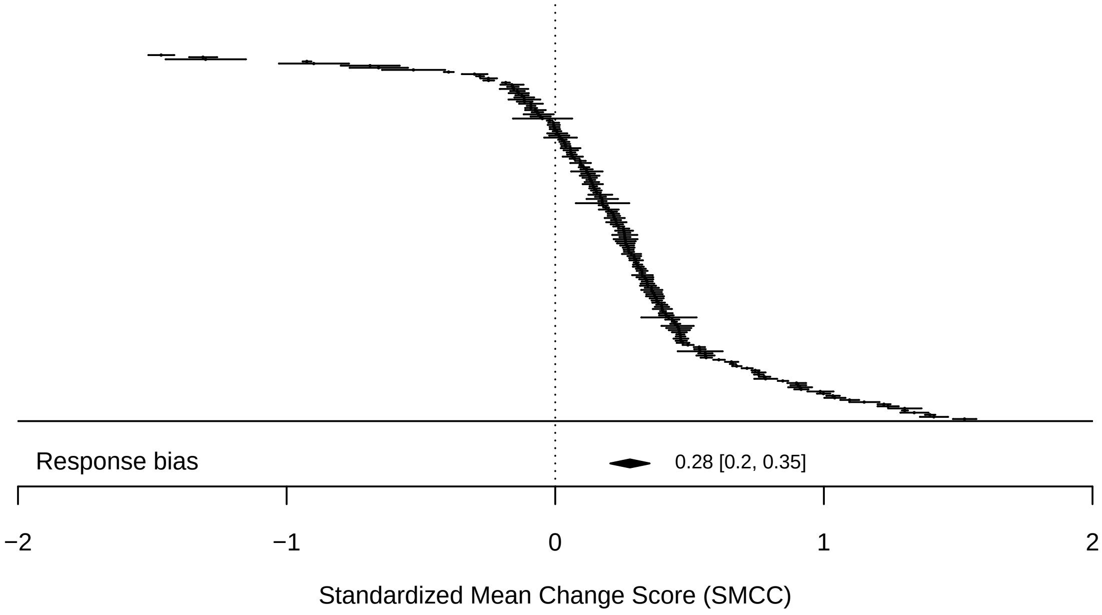

Figure 4. Forest plot for response bias. Effects are weighed by their sample size. Horizontal bars represent 95% confidence intervals. The average estimate is the result of a multilevel meta model with clustered standard errors at the sample level.

Discernment = 0.35 [0.14, 0.56]; baseline discernment other countries pooled = 0.83 [0.67, 0.98]). However, we did not find a statistically significant difference regarding response bias.

Scales. The studies in our meta analysis used a variety of response scales, including both binary (e.g. “Do you think the above headline is accurate? - Yes, No”) and continuous ones (e.g. “To the best of your knowledge, how accurate is the claim in the above headline” 1 = Not at all accurate, 4 = Very accurate).

Regarding discernment, the only scale that differed from the most common four point scale (Baseline discernment 4-point-scale = 1.11 [0.94, 1.29]) was the six point scale, yielding lower discernment (Δ Discernment = -0.41 [-0.7, -0.11]).

Regarding response bias, studies using a four point scale (Baseline response bias 4-point scale = 0.54 [0.39, 0.7]) reported a larger response bias compared to studies using a binary and a 7-point scale (Δ response bias = -0.40 [-0.59, -0.22] for binary scales; -0.46 [-0.67, -0.26] for 7-point scales).

Format. Studies using as stimuli headlines with pictures (Δ response bias = 0.25 [0.09, 0.42]; 43 effects), or headlines with pictures and a lede (Δ response bias = 0.24 [0.06, 0.41]; 41 effects), displayed a stronger response bias compared to studies relying on headlines with no picture/lede (Baseline response bias headlines only = 0.18 [0.09, 0.27]; 79 effects).

We do not find differences related to format for discernment.

Topic. We did not find any difference in discernment and response bias across news topic, when distinguishing between the categories “political”, “covid” and “other”.

Sources. In line with past findings, we did not observe any difference in discernment and response bias between studies displaying the source of the news items (73 effects) and those that did not (73 effects; for 27 this information was not explicitly provided).

Political Concordance. The moderators investigated above were (mostly) not experimentally manipulated within studies, but instead varied between studies, which impedes causal inference. Political concordance is an exception in this regard. It was manipulated within 17 different samples, across 7 different papers. In those experiments, typically, a pretest establishes the political slant of news headlines (e.g. pro-republican vs. pro-democrat). In the main study, participants then rate the accuracy for news items of both political slants, and provide information about their own political stance. The ratings of items are then grouped into concordant or discordant (e.g. pro-republican news rated by republicans will be coded as concordant while pro-republican news rated by democrats will be coded as discordant). Political concordance had no statistically significant effect on discernment. However, participants displayed a response bias only when rating politically discordant headlines (see Fig. 5). In particular, when rating concordant items, participants did not show a response bias (Baseline response bias concordant items = -0.10 [-0.36, 0.17]), while for discordant news items, participants displayed a positive response bias (Δ response bias = 0.60 [0.39, 0.81]). In other words, participants were not gullible when facing concordant news headlines (as would have suggested a negative response bias), but were skeptical when facing discordant ones.

Discussion

This meta-analysis sheds light on some of the most common fears voiced about false news. In particular, we investigated whether people are able to discern true from false news, and whether they are better at discriminating true news or false news (response bias). Across 173 effect sizes ($N_{participants}$ = 100777) from 30 countries across 6 continents, we found that people rated true news as much more accurate than false news ($d_{discernment}$ = 1.01 [0.9, 1.12]) and are slightly better at rating false news as inaccurate than at rating true news as accurate (d_{bias} = 0.28 [0.2, 0.35]).

The finding that people can discern true from false news when prompted to do so has important implications for interventions against misinformation. First, it suggests that most people do not lack the skills to spot false news. Instead, when they interact with false news, people may sometimes lack the motivation to use these skills or apply them selectively[33,34]. Thus, instead of teaching people how to spot false news, it may be more fruitful to design interventions targeting motivations, either by manipulating features of the environment in which people encounter news[35,36], or by intrinsically motivating people to use their skills and pay more attention to accuracy[33]. For instance, it has been shown that design features of current social media environments sometimes impede discernment[37]. Similarly, it has been suggested that interventions against misinformation should build on

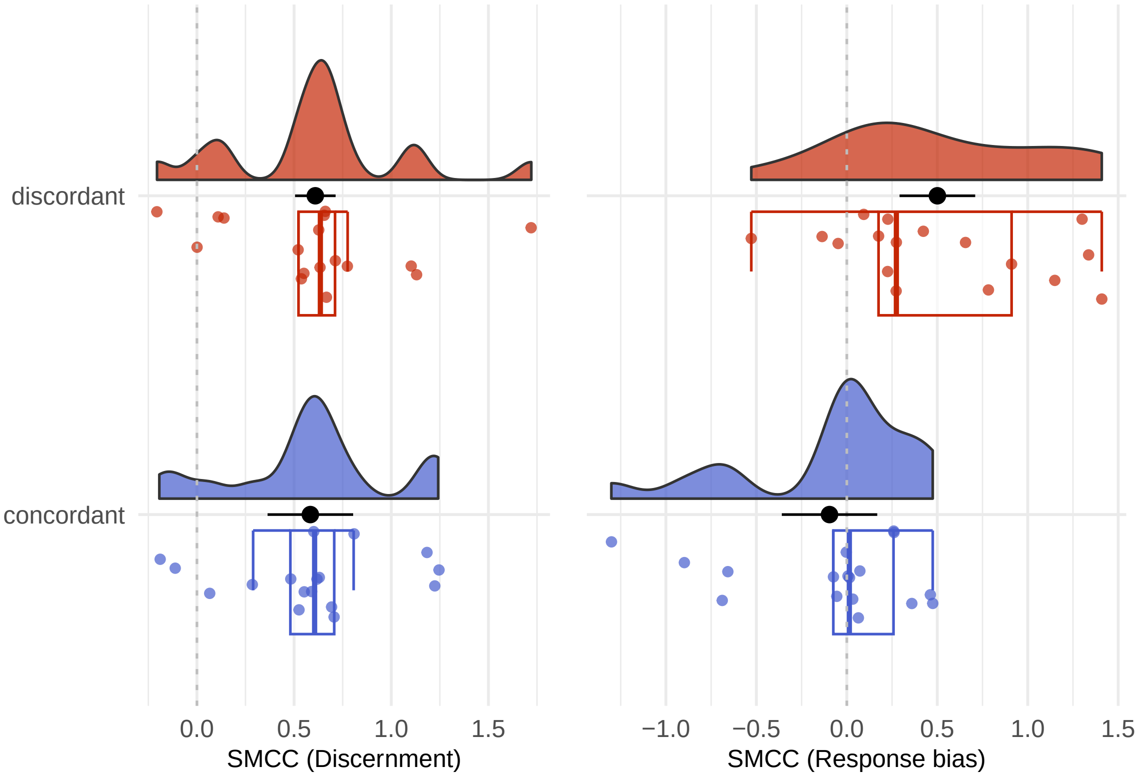

Figure 5. Distribution of effect sizes for politically concordant and discordant items. The black dots represent the predicted average of the meta-regression, the black horizontal bars the 95% confidence intervals.

the tacit knowledge that people rely on to discriminate false news, instead of giving people explicit tips and guidelines that people may struggle to integrate in their tacit knowledge[38].

Second, the fact that people can, on average, discern true from false news lends support to crowdsourced fact-checking initiatives. While fact-checkers cannot keep up with the pace of false news production, the crowd can, and it has been shown that even small groups of participants perform as well as professional fact-checkers[39,40]. The cross-cultural scope of our findings suggests that these initiatives may be fruitful in many countries across the world. Indeed, we found that in every country included in the meta-analysis, participants on average rated true news as more accurate than false news (see Appendix F).

The fact that people tend to disbelieve true news slightly more than they believe false news speaks to the nature of the misinformation problem and how to fight it. It is in line with the idea that the problem is less that people are gullible towards falsehoods, and more that they are skeptical towards reliable information[14,41]. Even assuming that the rejection of true news and the acceptance of false news are of similar scale (and that both can be improved), given that true news are much more prevalent in people's news diet than false news[42], true news skepticism may be more detrimental to the accuracy of people's beliefs than false news acceptance[13]. This skepticism is concerning in the context of the low and declining trust in the news across the world[43], the attacks of populist leaders on the news media[21], and growing news avoidance[43]. Interventions aimed at reducing misperceptions should therefore consider increasing the acceptance of true news in addition to reducing the acceptance of false news[13]. At the very least, when testing interventions, researchers should evaluate their effect on both true and false news, not just false news[44]. This is all the more important given that recent evidence suggest that many interventions against misinformation, such as media literacy tips[45], fact-checking[46], or educational games aimed at inoculating people against misinformation[27], may reduce misperceptions of false news at the cost of also reducing trust in true news.

We also investigated various moderators of discernment and response bias. We found that discernment was greater in studies conducted in the United States compared to the rest of the world. This could be due to the inclusion of many countries from the Global South, where belief in misinformation and conspiracy theories has been documented to be higher[47]. In line with past work[28], the presence of a source had no statistically significant effects. The topic of the news also had no statistically significant effects on discernment and response bias. Participants showed greater skepticism (higher response bias) in studies that presented headlines in a social media format (with an image and lede) or along with an image compared to studies that used plain headlines. This suggests that the skepticism of true news documented in this meta-analysis may be partially due to the social media format of the news headlines. Past work has shown that people report trusting news on social media less[2,18], and experimental manipulations have shown that the Facebook news format reduces belief in news[48,49] –although the causal effects documented in these experiments are much smaller than observational differences in reported trust levels between news on social media and on news outlets[50]. Lower trust of news on social media may be a good thing, given that on average news on social media may be less accurate than news on news websites, but is also worrying given that most of news consumption worldwide is shifting online and on social media in particular[43].

Finally, the political concordance of the news had no effect on discernment, but participants were excessively skeptical of politically discordant news. That is, participants were equally skilled at discerning true from false news for concordant and discordant items, but they rated news generally (true and false) as more false when politically discordant. This finding is in line with the idea that people are not excessively gullible of news they agree with, but are instead excessively skeptical of news they disagree with[14,51]. It suggests that interventions aimed at reducing partisan motivated reasoning, or at improving political reasoning in general, should focus more on increasing openness to opposing viewpoints than on increasing skepticism towards concordant viewpoints.

Our meta-analysis has a number of conceptual limitations. First, in all studies participants had to rate the accuracy of the news stories, which may have increased discernment, given that prompting people to think about accuracy increases sharing discernment[33]. In the wild, when browsing on social media, people may be less discerning than in these experimental settings because they would pay less attention to accuracy[37]. However, given people's low exposure to misinformation online[52], most people may protect themselves from misinformation not by detecting misinformation on the spot, but by relying on the reputation of the sources and avoiding unreliable sources[53]. Second, accuracy ratings were averaged across participants and thus better reflect the wisdom of the crowd than the skills of individuals. Yet, past work[39] shows that most individuals appear able to discern true from false news better than chance. In line with this, studies for which we have raw data show that 87.5% of individual participants rated true news as more accurate than false news (see Appendix C). Third, the vast majority of studies in our meta-analysis relied on fact-checked false news. It is unclear whether the present findings generalize to non fact-checked false news and misinformation more broadly. For instance, it is likely that discerning true from misleading news is harder than discerning true from false news, and that people may discriminate true news better than they discriminate misleading news[54]. Fourth, response bias could be an artifact of biased news selection for experiments. For example, researchers may have selected true news that are not obviously true to avoid ceiling effects. However, we deem this unlikely and address it in Appendix G, notably by showing that studies which randomly sampled a large number of headlines from high-quality mainstream news sites report accuracy ratings of true news similar to the average true news rating across studies in our meta analysis.

Our meta-analysis further has methodological limitations which we address in a series of robustness checks in the appendix. We show that our results hold across a wide range of imputed correlation values for our effect size estimation and that the resulting estimate is conservative compared to alternative effect size estimators (Appendix A). We also show that we obtain similar results when running a participant-level analysis on a subset of studies for which we have raw data (Appendix C) and when collapsing Likert scales into binary ones for that subset (Appendix D).

In conclusion, we found that in experimental settings, people are able to discern true news from false news, but when they err, they tend to do so on the side of skepticism more than on the side of gullibility. These findings lend support to crowdsourced fact-checking initiatives, and suggest that, to improve discernment, there may be more room to increase the acceptance of true news than to reduce the acceptance of false news.

Methods

Data

We undertook a systematic review and meta-analysis of the experimental literature on accuracy judgements of news, following the PRISMA guidelines[55]. All records resulting from our literature searches can be found on the OSF project page. We documented rejection decisions for all retrieved papers. They, too, can be found on the OSF project page.

Deviations from eligibility criteria. We followed our eligibility criteria (as outlined above), with 4 exceptions. We rejected one paper based on a criterion that we had not previously set: scale asymmetry.[56] asked participants: "According to your knowledge, how do you rate the following headline?", providing a very asymmetrical set of answer options ("1—not credible; 2—somehow credible; 3—quite credible; 4—credible; 5—very credible"). The paper provides 6 effect sizes, all of which strongly favor our second hypothesis (one effect being as large as d = 2.54). We decided to exclude this paper from our analysis because of its very asymmetric scale. Further, we stretched our criterion for real-world news on three instances.[57] and[58] used artificial intelligence trained on real-world news to

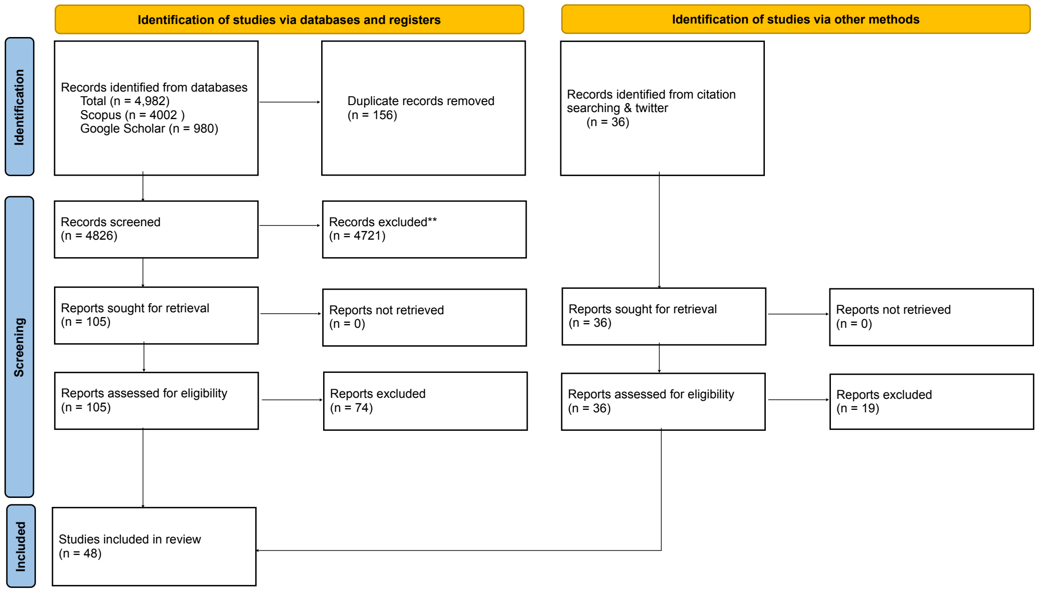
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from databases
Total (n = 4,982)
Scopus (n = 4002)
Google Scholar (n = 980)
Duplicate records removed
(n = 156)
Records identified from citation searching & twitter
(n = 36)
Records screened
(n = 4826)
Records excluded**
(n = 4721)
Reports sought for retrieval
(n = 105)
Reports not retrieved
(n = 0)
Reports sought for retrieval
(n = 36)
Reports not retrieved
(n = 0)
Reports assessed for eligibility
(n = 105)
Reports excluded
(n = 74)
Reports assessed for eligibility
(n = 36)
Reports excluded
(n = 19)
Studies included in review
(n = 48)

Figure 6. 2020 PRISMA flow diagram for new systematic reviews.

generate false news.[59] had journalists create the false news items. We reasoned that asking journalists to write news should be similar enough to real-wolrd news, and that LLMs already produce news headlines that are indistinguishable from real news, so it should not make a big difference.

Literature search. We first conducted a Scopus search (search string: *‘ “false news” OR “fake news” OR “false stor*” AND “accuracy” OR “discernment” OR “credibilit*” OR “belief” OR “susceptib*” ’*).

Given the high volume of papers (12425), we added restrictions to only include articles that were likely (i) experimental, (ii) and exposed participants to both true and false news (addition to search string: *‘AND (LIMIT-TO (LANGUAGE , “English”)) AND (LIMIT-TO (DOCTYPE , “ar”) OR LIMIT-TO (DOCTYPE , “cp”)) AND (EXCLUDE (SUBJAREA , “PHYS”) OR EXCLUDE (SUBJAREA , “MATE”) OR EXCLUDE (SUBJAREA , “BIOC”) OR EXCLUDE (SUBJAREA , “ENER”) OR EXCLUDE (SUBJAREA , “IMMU”) OR EXCLUDE (SUBJAREA , “AGRI”) OR EXCLUDE (SUBJAREA , “PHAR”) OR EXCLUDE (SUBJAREA , “HEAL”) OR EXCLUDE (SUBJAREA , “EART”) OR EXCLUDE (SUBJAREA , “NURS”) OR EXCLUDE (SUBJAREA , “CHEM”) OR EXCLUDE (SUBJAREA , “CENG”) OR EXCLUDE (SUBJAREA , “VETE”) OR EXCLUDE (SUBJAREA , “DENT”)) AND (EXCLUDE (SUBJAREA , “COMP”) OR EXCLUDE (SUBJAREA , “ENGI”) OR EXCLUDE (*

SUBJAREA , "MATH") OR EXCLUDE (SUBJAREA , "MEDI")').

We excluded papers not written in English, that were not articles or conference papers, and that were from disciplines that are likely irrelevant for the present search (e.g., Dentistry, Veterinary, Chemical Engineering, Chemistry, Nursing, Pharmacology, Microbiology, Materials Science, Medicine) or unlikely to use an experimental design (e.g. Computer Science, Engineering, Mathematics). After these filters were applied, we ended up with 4002 results. Second, we conducted an advanced Google Scholar search via the "Publish or Perish" software (search string: *' "Fake news" | "False news"|"False stor*" "Accuracy" | "Discernment"|"Credibility"|"Belief"|"Suceptib*", no citations, no patents'*). The main advantage of this search was to identify important pre-prints or working papers that the Scopus search would have missed. The Google scholar search yielded 980 results. After removing 156 duplicates from the two searches, we ended up with 4826 documents for screening. We screened records based on titles only. The vast majority of documents (4721) had irrelevant titles and were removed during that phase. Most irrelevant titles were not about false news or misinformation (e.g. "Formation of a tourist destination image: Co-occurrence analysis of destination promotion videos"), and some were about false news or misinformation but were not about belief or accuracy (e.g. "Freedom of Expression and Misinformation Laws During the COVID-19 Pandemic and the European Court of Human Rights"). We stored the remaining 105 records in the reference management system Zotero for retrieval. Of those, we rejected a total of 74 papers that did not meet our inclusion criteria. We rejected 5 papers based on their abstract and 69 after assessment of the full text. We included the remaining 31 papers from the systematic literature search. To complement the systematic search results, we conducted forward and backward citation search through Google Scholar. We also reviewed additional studies that we had on our computers and papers we found scrolling through twitter (mostly unpublished manuscripts). Taken together we identified an additional 36 papers via those methods. Of these, we excluded 19 papers after full text assessment because they did not meet our inclusion criteria. For these papers, too, we documented our exclusion decisions. They can be found together with the ones of the systematic search on the OSF project page. We included the remaining 17 papers. In total, we included 48 papers in our meta analysis, 34 of which were peer-reviewed and 14 grey literature (reports and working papers). We retrieved the relevant summary statistics directly from the paper for 20 papers, calculated them ourselves based on publicly available raw data for 16 papers, and got them from the authors after request for 12 papers.

Statistical methods

All data and code are publicly available on the OSF project page. Unless explicitly mentioned otherwise, we pre-registered all reported analyses. Our choice of statistical models was informed by simulations, which can also be found on the OSF project page. We conducted all analyses in R[60] using Rstudio[61] and the `tidyverse` package[62]. For effect size calculations, we rely on the `escalc()`, for for models on the `rma.mv()`, for clustered standard errors on the `robust()` function, all from the `metafor` package[63].

Deviations from pre-registration. We did not pre-register considering scale symmetry and proportion of true news as a moderator variable. We report the results regarding these variables in Appendix B.

Outcomes. We have two complementary measures of assessing the quality of people's news judgment. The first measure is discernment. It measures the overall quality of news judgment across true and false news. We calculate discernment by subtracting the mean accuracy ratings of false news from the mean accuracy ratings of true news, such that more positive scores indicate better discernment. However, discernment is a limited diagnostic of the quality of people's news judgment. Imagine a study A in which participants rate 50% of true news and 20% of false news as accurate, and a study B finding 80% of true news and 50% of false news rated as accurate. In both cases, the discernment is the same: Participants rated true news as more accurate by 30 percentage points than false news. However, the performance by news type is very different. In study A, people do well for false news - they only mistakenly classify 20% as accurate - but are at chance for true news. In study B, it's the opposite. We therefore use a second measure: response bias. For any given level of discernment, it indicates whether people's judgements were better on true news or on false news, and to what extent. First, we calculate an error for false and true news separately, which we define as the distance of participants' actual ratings to the best possible ratings. For example, for study A, the mean error for true news is 50% (100%-50%), because in the best possible scenario, participants would have classified 100% of true news as true. The error for false news in Study A is 20% (20%-0%), because the best possible performance for participants would have been to classify 0% of false news as accurate. We calculate response bias by subtracting the mean error for false news from the mean error for true news. For example, for Study A, the response bias is 30% (50%-20%). A positive response bias indicates that people doubt true news more than they believe false news.

Effect sizes. The studies in our meta analysis used a variety of response scales, including both binary (e.g. "Do you think the above headline is accurate? - Yes, No") and continuous ones (e.g. "To the best of your knowledge, how accurate is the claim in the above headline" 1 = Not at all accurate, 4 = Very accurate). To be able to compare across the different scales, we calculated standardized effects, i.e. effects expressed in units of standard deviations. Common standardized mean difference (SMD) measures such as Cohen's d or Hedge's g assume that groups (in our case false and true news ratings) are independent. However, the vast majority of experiments (168 out of 173 effects) in our meta analysis manipulated news veracity within participants, i.e. having participants rate both false and true news. To account for the dependency between ratings that this design generates, we calculated standardized mean changes using change score standardization (SMCC)[31]. This change score is calculated as

$$SMCC = \frac{MD}{SD_d}$$

with MD being the mean difference/change score (mean true news score minus mean false news score) and SD_d being standard deviation of the difference/change scores, which (assuming equal standard deviations for false and true news) is calculated as: $SD_d = SD_{false/true}\sqrt{2(1-r)}$[64].

Similar to more common standardized effect measures such as Cohen's d, the SMCC is a measure of mean difference in terms of a pooled standard deviation. The difference is that the SMCC does not assume independence between groups, by taking into account the correlation between false and true news when calculating the pooled standard deviation. The SMCC varies on the imputed correlation value r, because SD_d varies as a function of r. If r is greater than .5, SD_d will be smaller than $SD_{false/true}$, and as a result, the SMCC will be larger than the estimate obtained by a standardized mean difference assuming independence such as Cohen's d. By contrast, when the correlation is less than .5, SD_d will be greater than $SD_{false/true}$, and the SMCC will be smaller[64]. Ideally, for each effect size (i.e. the meta-analytic units of observation) in our data, we need an estimate of the correlation between false and true news ratings. However, this correlation is generally not reported in the original paper. We could only obtain it for a subset of samples for which we collected the summary statistics ourselves, based on the raw data. Based on this subset of correlations, we calculated an average correlation, which we then imputed for all effect size calculations. This approach is in line with the Cochrane recommendations for crossover trials[65]. In our case, this average correlation is 0.16. In appendix A we run sensitivity analyses which show that our results hold across all correlation values that we obtain from individual-level data. For the 5 (out of 173) effects from studies that used a between participant design, we calculated Hedge's g[32]. For all effect size calculations, we defined the sample size as the number of instances of news ratings. That is, we multiplied the number of participants with the number of news items rated per participant.

Models. In our models for the meta analysis, each effect size was weighted by the inverse of its standard error, thereby giving more weight to studies with larger sample sizes. We used random effects models, which assume that there is not only one true effect size but a distribution of true effect sizes[66]. These models assume that variation in effect sizes is not only due to sampling error alone, and thereby allow to model other sources of variance. We estimated the overall effect of our outcome variables using a three-level meta-analytic model with random effects on the sample and the publication level. This approach allowed us to account for the hierarchical structure of our data, in which samples (level three) contribute multiple effects (level two)[5]. Multiple effects per sample occur, for example, when separate accuracy ratings are available by news topic, or when follow-up studies were conducted on the same participants. However, the multi-level models do not account for dependencies in sampling error. When one same sample contributes several effect sizes, one should expect their respective sampling errors to be correlated[66]. To account for dependency in sampling errors, we computed cluster-robust standard errors, confidence intervals, and statistical tests for all estimated effect sizes.

To assess the effect of moderator variables, we calculated meta regressions. We calculated a separate regression for each moderator, by adding the moderator variable as a fixed effect to the multilevel meta models presented above. We pre-registered a list of six moderator variables to test. Those included the *country* of studies (levels: United States vs. all other countries), *political concordance* (levels: politically concordant vs. politically discordant), *news family* (levels: political, including both concordant and discordant vs. covid related vs. other, including categories as diverse as history, environment, health, science

[5] Level 1 being the participant level of the original studies, see[66].

and military related news items), the *format* in which the news were presented (levels: headline only vs. headline and picture vs. headline, picture and lede), whether news items were accompanied by a *source* or not, and the *response scale* used (levels: 4-point vs. binary vs. 6-point vs. 7-point vs. other, for all other numeric scales that were not frequent). We ran an additional regression on a non-preregistered variable, namely the *symmetry of scales* (levels: perfectly symmetrical vs. imperfectly symmetrical). We further descriptively checked whether the *proportion of true news* among all news would yield differences.

Publication bias. We ran some standard procedures for detecting publication bias. However, a priori we did not expect publication bias to be present because our variables of interest were not those of interest to the researchers of the original studies: Researchers generally set out to test factors that alter discernment, and not the state of discernment in the control group. No study measured response bias in the way we define it here.

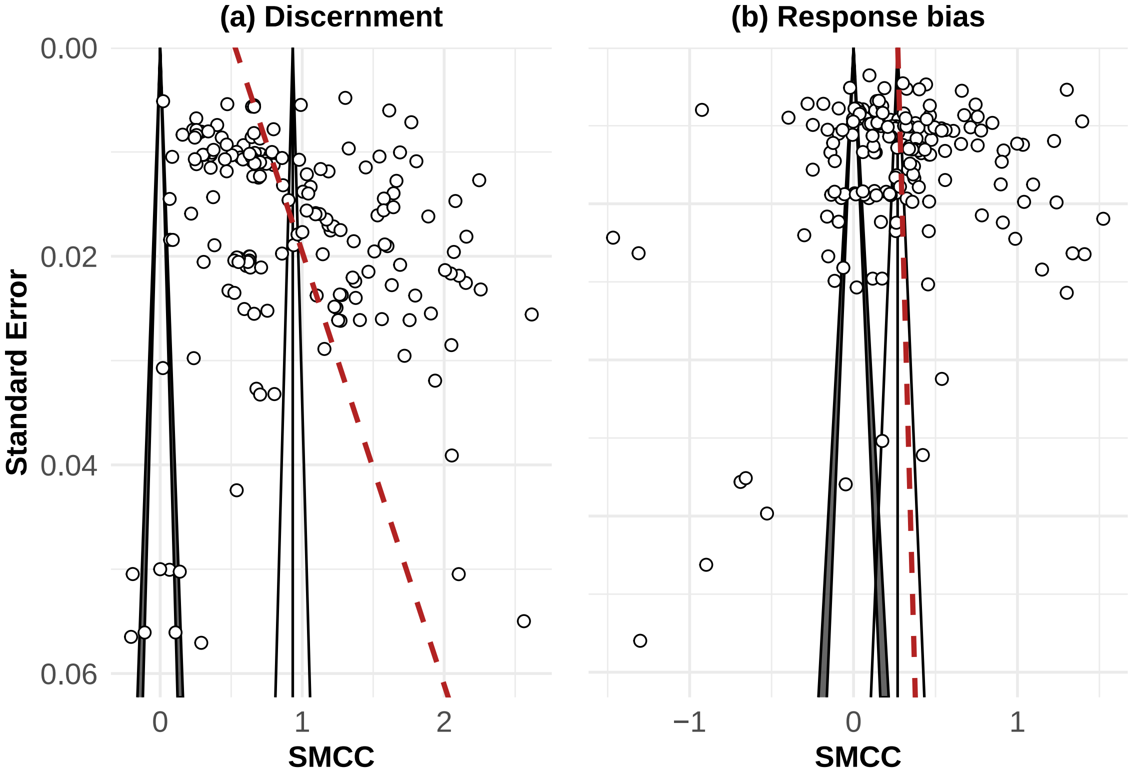

Figure 7. Funnel plots for discernment and response bias. Dots represent effect sizes. In the absence of publication bias and heterogeneity, one would then expect to see the points forming a funnel shape, with the majority of the points falling inside of the pseudo-confidence region centered around the average effect estimate, with bounds of ±1.96 SE (the standard error value from the y-axis). The dashed red regression line illustrates the estimate of the Egger's regression test. For discernment, its slope differs significantly from zero, suggesting that smaller studies tend to report larger effect sizes.

Regarding discernment, we find evidence that smaller studies tend to report larger effect sizes, according to Egger's regression test (see Fig. 7; see also Appendix E). However, it is unclear how meaningful this difference is. As illustrated by the funnel plot, there is generally high between-effect size heterogeneity: Even when focusing only on the most precise effect sizes (top of the funnel), the estimates vary substantially. It thus seems reasonable to assume that most of the dispersion of effect sizes does not arise from studies' sampling error, but from studies estimating different true effects. Further, even the small studies are relatively high powered, suggesting that they would have yielded significant, publishable results even with smaller effect sizes. Lastly, Egger's regression test can lead to an inflation of false positive results when applied to standardized mean differences[66,67]. We do not find evidence for asymmetry regarding response bias.

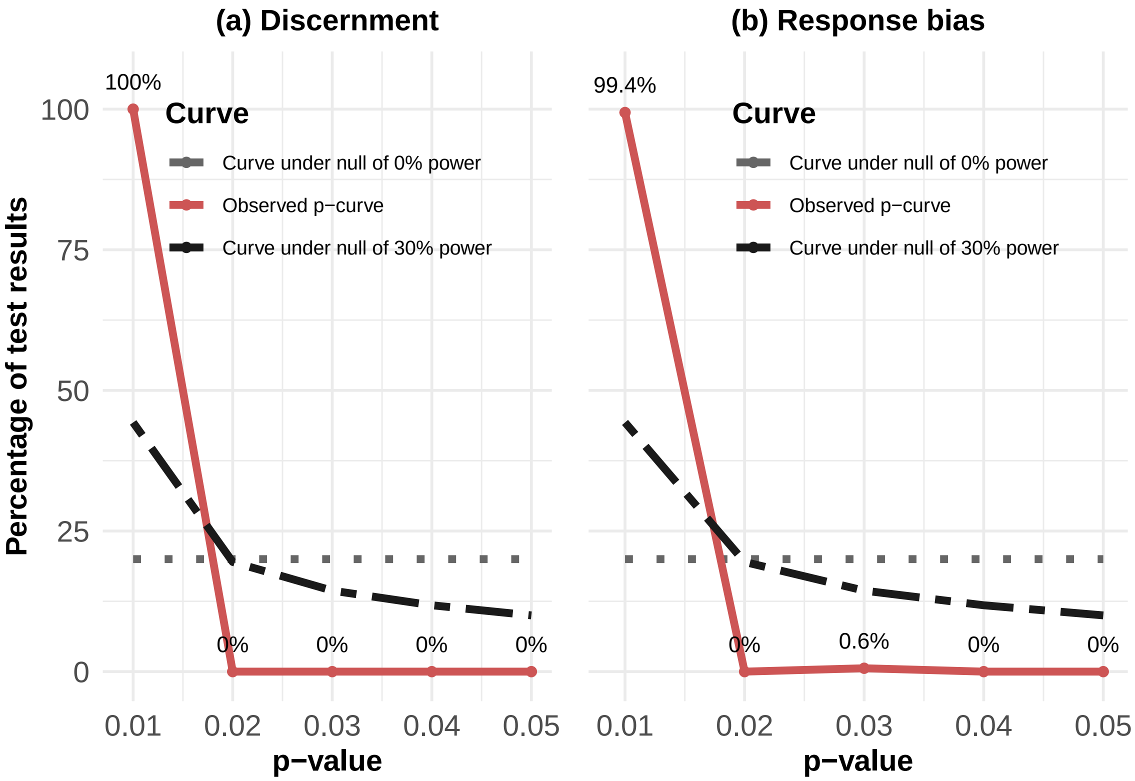

Figure 8. P-curves for discernment and response bias. The p-curve shows the percentage of of effect sizes for a given p value within the range of 0.1 and 0.5. All values smaller than 0.01 are rounded to that value. The reference lines indicate the expected percentage of studies for a given p value, assuming that there is a true effect and certain statistical power to detect it (either 0% or 30% power). The observed p-curve is negatively sloped and heavily right skewed (the tail points to the right) for both outcomes, which suggests no widespread p-hacking.

We do not find any evidence to suspect p-hacking for either discernment or response bias from visually inspecting p-curves for both outcomes (see Fig. 8).

Data availability. The extracted data used to produce our results are available on the OSF project page (https://osf.io/96zbp/).

Code availability. The code used to create all results (including tables and figures) of this manuscript is also available on the OSF project page (https://osf.io/96zbp/).

References

1. Brennen, T. & Magnussen, S. Lie Detection: What Works? *Current Directions in Psychological Science* 096372142311730 (2023) doi:10.1177/09637214231173095.

2. Mont'Alverne, C. *et al.* The trust gap: How and why news on digital platforms is viewed more sceptically versus news in general. (2022).

3. Metzger, M. J. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology* **58**, 2078–2091 (2007).

4. Ross Arguedas, A. *et al.* Snap judgements: How audiences who lack trust in news navigate information on digital platforms. (2022).

5. Altay, S., Lyons, B. & Modirrousta-Galian, A. Exposure to higher rates of false news erodes media trust and fuels skepticism in news judgment. doi:10.31234/osf.io/t9r43.

6. Bryanov, K. & Vziatysheva, V. Determinants of individuals' belief in fake news: A scoping review determinants of belief in fake news. *PLOS ONE* **16**, e0253717 (2021).

7. Pennycook, G. & Rand, D. G. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* **188**, 39–50 (2019).

8. Schulz, A., Fletcher, R. & Popescu, M. Are news outlets viewed in the same way by experts and the public? A comparison across 23 european countries. *Reuters Institute for the Study of Journalism* (2020).

9. Mourão, R. R. & Robertson, C. T. Fake News as Discursive Integration: An Analysis of Sites That Publish False, Misleading, Hyperpartisan and Sensational Information. *Journalism Studies* **20**, 2077–2095 (2019).

10. Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146–1151 (2018).

11. Chen, X., Pennycook, G. & Rand, D. What makes news sharable on social media? *Journal of Quantitative Description: Digital Media* **3**, (2023).

12. Altay, S., Araujo, E. de & Mercier, H. "If This account is True, It is Most Enormously Wonderful": Interestingness-If-True and the Sharing of True and False News. *Digital Journalism* **10**, 373–394 (2022).

13. Acerbi, A., Altay, S. & Mercier, H. Research note: Fighting misinformation or fighting for information? *Harvard Kennedy School Misinformation Review* (2022) doi:10.37016/mr-2020-87.

14. Mercier, H. *Not born yesterday: the science of who we trust and what we believe.* (2020).

15. Brashier, N. M. & Marsh, E. J. Judging Truth. *Annual Review of Psychology* **71**, 499–515 (2020).

16. Street, C. N. H. & Masip, J. The source of the truth bias: Heuristic processing? *Scandinavian Journal of Psychology* **56**, 254–263 (2015).

17. Levine, T. R. Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology* **33**, 378–392 (2014).

18. Newman, N., Fletcher, R., Robertson, C. T., Eddy, K. & Nielsen, R. K. Reuters Institute Digital News Report 2022. (2022).

19. Mihailidis, P. & Foster, B. The Cost of Disbelief: Fracturing News Ecosystems in an Age of Rampant Media Cynicism. *American Behavioral Scientist* **65**, 616–631 (2021).

20. Egelhofer, J. L. & Lecheler, S. Fake news as a two-dimensional phenomenon: a framework and research agenda. *Annals of the International Communication Association* **43**, 97–116 (2019).

21. Van Duyn, E. & Collier, J. Priming and Fake News: The Effects of Elite Discourse on Evaluations of News Media. *Mass Communication and Society* **22**, 29–48 (2019).

22. Paul, C. & Matthews, M. The russian "firehose of falsehood" propaganda model. *Rand Corporation* **2**, 1–10 (2016).

23. Ulusoy, E. *et al.* Flooding the zone: How exposure to implausible statements shapes subsequent belief judgments. *International Journal of Public Opinion Research* **33**, 856–872 (2021).

24. Fletcher, R. & Nielsen, R.-K. People dont trust news media–and this is key to the global misinformation debate. *AA. VV., Understanding and Addressing the Disinformation Ecosystem* 13–17 (2017).

25. Luo, M., Hancock, J. T. & Markowitz, D. M. Credibility Perceptions and Detection Accuracy of Fake News Headlines on Social Media: Effects of Truth-Bias and Endorsement Cues. *Communication Research* **49**, 171–195 (2022).

26. Batailler, C., Brannon, S. M., Teas, P. E. & Gawronski, B. A Signal Detection Approach to Understanding the Identification of Fake News. *Perspectives on Psychological Science* **17**, 78–98 (2022).

27. Modirrousta-Galian, A. & Higham, P. A. Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with receiver operating characteristic analysis. *Journal of Experimental Psychology: General* (2023).

28. Dias, N., Pennycook, G. & Rand, D. G. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* (2020) doi:10.37016/mr-2020-001.

29. Tappin, B. M., Pennycook, G. & Rand, D. G. Bayesian or biased? Analytic thinking and political belief updating. *Cognition* **204**, 104375 (2020).

30. Calvillo, D. P. & Smelter, T. J. An initial accuracy focus reduces the effect of prior exposure on perceived accuracy of news headlines. *Cognitive Research: Principles and Implications* **5**, 55 (2020).

31. Gibbons, R. D., Hedeker, D. R. & Davis, J. M. Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics* **18**, 271–279 (1993).

32. Hedges, L. V. Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics* **6**, 107–128 (1981).

33. Pennycook, G. *et al.* Shifting attention to accuracy can reduce misinformation online. *Nature* **592**, 590–595 (2021).

34. Rathje, S., Roozenbeek, J., Van Bavel, J. J. & Van Der Linden, S. Accuracy and social motivations shape judgements of (mis)information. *Nature Human Behaviour* (2023) doi:10.1038/s41562-023-01540-w.

35. Capraro, V. & Celadin, T. "I Think This News Is Accurate": Endorsing Accuracy Decreases the Sharing of Fake News and Increases the Sharing of Real News. *Personality and Social Psychology Bulletin.*

36. Globig, L. K., Holtz, N. & Sharot, T. Changing the incentive structure of social media platforms to halt the spread of misinformation. *eLife* **12**, e85767 (2023).

37. Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G. & Rand, D. The social media context interferes with truth discernment. *Science Advances* **9**, eabo6169 (2023).

38. Modirrousta-Galian, A., Higham, P. A. & Seabrooke, T. *Wordless Wisdom: The Dominant Role of Tacit Knowledge in True and Fake News Discrimination.* https://osf.io/2gubk (2023) doi:10.31234/osf.io/2gubk.

39. Allen, J., Arechar, A. A., Pennycook, G. & Rand, D. G. Scaling up fact-checking using the wisdom of crowds. *Science advances* **7**, eabf4393 (2021).

40. Martel, C., Allen, J. N. L., Pennycook, G. & Rand, D. G. *Crowds Can Effectively Identify Misinformation at Scale.* https://osf.io/2tjk7 (2022) doi:10.31234/osf.io/2tjk7.

41. Altay, S., Berriche, M. & Acerbi, A. Misinformation on Misinformation: Conceptual and Methodological Challenges. *Social Media.*

42. Allen, J., Howland, B., Mobius, M., Rothschild, D. & Watts, D. J. Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances* **6**, eaay3539 (2020).

43. Newman, N., Fletcher, R., Eddy, K., Robertson, C. T. & Nielsen, R. K. Digital news report 2023. (2023).

44. Guay, B., Berinsky, A., Pennycook, G. & Rand, D. How to think about whether misinformation interventions work. doi:10.31234/osf.io/gv8qx.

45. Hoes, E., Aitken, B., Zhang, J., Gackowski, T. & Wojcieszak, M. *Prominent misinformation interventions reduce misperceptions but increase skepticism.* https://osf.io/zmpdu (2023) doi:10.31234/osf.io/zmpdu.

46. Bachmann, I. & Valenzuela, S. Studying the Downstream Effects of Fact-Checking on Social Media: Experiments on Correction Formats, Belief Accuracy, and Media Trust. *Social Media + Society* **9**, 20563051231179694 (2023).

47. Alper, S. When conspiracy theories make sense: The role of social inclusiveness.

48. Besalú, R. & Pont-Sorribes, C. Credibility of Digital Political News in Spain: Comparison between Traditional Media and Social Media. *Social Sciences* **10**, 170 (2021).

49. Karlsen, R. & Aalberg, T. Social Media and Trust in News: An Experimental Study of the Effect of Facebook on News Story Credibility. *Digital Journalism* **11**, 144–160 (2023).

50. Agadjanian, A. *et al.* A platform penalty for news? How social media context can alter information credibility online. *Journal of Information Technology & Politics* **20**, 338–348 (2023).

51. Trouche, E., Johansson, P., Hall, L. & Mercier, H. Vigilant conservatism in evaluating communicated information. *PLOS ONE* **13**, e0188825 (2018).

52. Altay, S., Kleis Nielsen, R. & Fletcher, R. Quantifying the "infodemic": People turned to trustworthy news outlets during the 2020 coronavirus pandemic. *Journal of Quantitative Description: Digital Media* **2**, (2022).

53. Altay, S., Hacquin, A.-S. & Mercier, H. Why do so few people share fake news? It hurts their reputation. *new media* 22.

54. Aslett, K. *et al.* An Ecologically and Externally Valid Approach to Assessing Belief in Popular Misinformation.

55. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).

56. Baptista, J. P., Correia, E., Gradim, A. & Piñeiro-Naval, V. The Influence of Political Ideology on Fake News Belief: The Portuguese Case. *Publications* **9**, 23 (2021).

57. Maertens, R. *et al.* *The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment.* https://osf.io/gk68h (2021) doi:10.31234/osf.io/gk68h.

58. Roozenbeek, J. *et al.* Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science* **7**, 201199 (2020).

59. Bryanov, K. *et al.* What Drives Perceptions of Foreign News Coverage Credibility? A Cross-National Experiment Including Kazakhstan, Russia, and Ukraine. *Political Communication* **40**, 115–146 (2023).

60. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, 2022).

61. Posit team. *RStudio: Integrated development environment for r.* (Posit Software, PBC, 2023).

62. Wickham, H. *et al.* Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).

63. Viechtbauer, W. Conducting meta-analyses in *r* with the **metafor** package. *J. Stat. Soft.* **36**, (2010).

64. Morris, S. B. & DeShon, R. P. Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods* **7**, 105–125 (2002).

65. Higgins, J. P. *et al.* *Cochrane handbook for systematic reviews of interventions.* (John Wiley & Sons, 2019).

66. Harrer, M., Cuijpers, P., A, F. T. & Ebert, D. D. *Doing meta-analysis with r: A hands-on guide.* (Chapman & Hall/CRC Press, 2021).

67. Pustejovsky, J. E. Simulating correlated standardized mean differences for meta-analysis. (2019).

68. Becker, B. J. Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology* **41**, 257–278 (1988).

69. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software* **67**, (2015).

70. Egger, M., Smith, G. D., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**, 629–634 (1997).

71. Stewart, A. J., Arechar, A. A., Rand, D. G. & Plotkin, J. B. The distorting effects of producer strategies: Why engagement does not reliably reveal consumer preferences for misinformation. doi:10.48550/arXiv.2108.13687.

72. Clemm Von Hohenberg, B. Truth and Bias, Left and Right: Testing Ideological Asymmetries with a Realistic News Supply. *Public Opinion Quarterly* **87**, 267–292 (2023).

73. Aslett, K. *et al.* An Ecologically and Externally Valid Approach to Assessing Belief in Popular Misinformation. (2023).

Appendix A
Effect sizes

Alternative effect sizes

Table A1 shows the estimated SMCC for both discernment (H1) and response bias (H2). For reference, we included the estimates for two alternative estimators: A standardized mean difference assuming independence (SMD), precisely Hedge's g, and a standardized mean change using raw (instead of change) score standardization (SMCR)[68]. When using raw score standardization, the standardized mean change expresses the effect size in terms of the standard deviation units of the pre-treatment (in our case false news) scores, rather than the standard deviation of the difference scores (involving the correlation)[68]. While our SMCC depends on the value of the correlation between the false and true news scores, the SMD and the SMCR do not. The interpretation of all three effect measures is similar: all are expressed in terms of standard deviations. Yet, they are different estimators, because they rely on different standard deviations, thereby producing different estimates and standard errors[64]. Due to the low average correlation between false and true news ratings, the SMCC produces the most conservative (i.e. smallest) effect estimates for both discernment and response bias.

Sensitivity analysis for imputed correlation

When using *standardized* effect measures, accounting for dependency by imputing correlations impacts the magnitude of the effect estimate itself. That is because, in this case, the standard deviation is not only used to calculate the standard error (as with *non* standardized effects), but also for the effect itself. Standardized effect measures depending on imputed correlations should therefore be accompanied by a sensitivity analysis, i.e. check how sensitive the effect estimate is to different imputed correlation values[65]. Figures A1 and

Table A1
Model results

	Main estimator (preregistered)		*Alternative estimators*			
	SMCC		SMCR		SMD	
	Discernment	Response bias	Discernment	Response bias	Discernment	Response bias
Estimate	1.012***	0.278***	1.334***	0.365***	1.305***	0.365***
	(0.055)	(0.037)	(0.076)	(0.047)	(0.071)	(0.047)
Num.Obs.	172	172	172	172	172	172
AIC	252.1	219.4	362.3	296.0	342.0	300.1
BIC	261.5	228.9	371.7	305.4	351.5	309.5

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Comparison of different effect sizes. The SMCC (Standardized mean change using change score standardization) is the estimator we pre-registered and report in the main analysis. For reference, we provide the results we obtain when using a standardized mean difference assuming independence (SMD), precisely Hedge's g, and a standardized change score using raw (instead of change) standardization (SMCR). For effects from studies that used a between participant design, we always calculated Hedge's G, including in those results listed under "SMCC" and "SMCR".

Table A2

	1-point	10-point	100-point	21-point	4-point	6-point	7-point	binary
Papers	3	2	1	1	17	8	8	13
Samples	25	3	1	1	37	17	29	20
Effects	25	3	1	2	45	27	43	27

Note. Frequency table of scales.

A2 show that our results hold when imputing any of the other correlation values occurring in the individual-level data. Moreover, all estimates from those imputations yield conservative (i.e. smaller) estimates compared to Hedges' g.

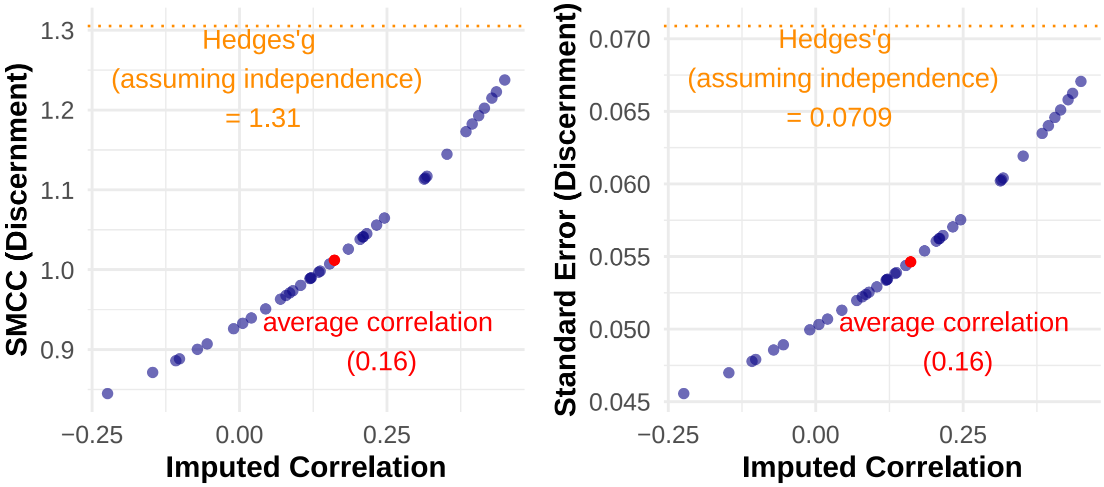

Figure A1. Estimated effect and standard error for discernment as a function of the imputed correlation value. Each dot represents an estimate corresponding to an imputed correlation value. There is one dot for each observed intra-sample correlation in the individual-level data. For reference, the horizontal dotted lines mark the estimate obtained when using Hedges'g, an effect size assuming independence.

Effects on original scales

Table A3 shows estimates by scale, in the original units of the scale. The table is intended to help interpret the magnitude of the effect sizes reported in the main findings. Note that some scales occur very rarely only (see Tab. A2), hence making their meta-analytic estimates less meaningful.

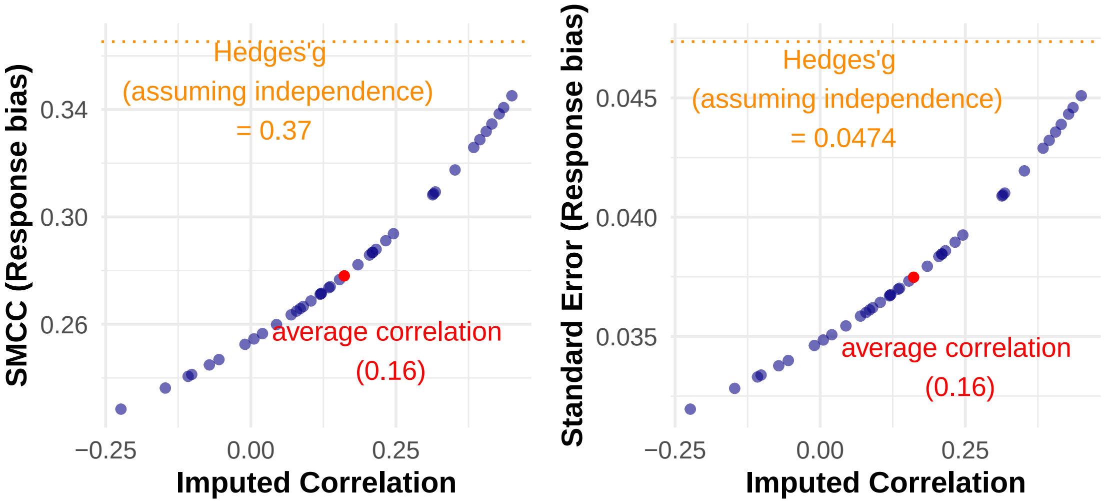

Figure A2. Estimated effect and standard error for response bias as a function of the imputed correlation value. Each dot represents an estimate corresponding to an imputed correlation value. There is one dot for each observed intra-sample correlation in the individual-level data. For reference, the horizontal dotted lines mark the estimate obtained when using Hedges'g, an effect size assuming independence.

Table A3

(Raw) Mean Differences between true and false news

	4-point	10-point	binary	7-point	6-point	1-point	21-point
Discernment							
Estimate	0.885***	2.440***	0.419***	1.600***	1.085***	0.290***	3.249***
	(0.060)	(0.170)	(0.030)	(0.201)	(0.081)	(0.023)	(0.414)
Num.Obs.	44	2	26	42	26	24	1
AIC	35.3	6.2	-40.7	112.1	24.7	-32.8	6.5
BIC	40.7	2.3	-36.9	117.3	28.5	-29.3	2.5
Response bias							
Estimate	0.403***	-1.807+	0.046	0.131	0.654***	0.092***	4.361***
	(0.054)	(1.078)	(0.029)	(0.095)	(0.150)	(0.017)	(0.858)
Num.Obs.	44	2	26	42	26	24	1
AIC	36.4	17.3	-30.3	76.6	68.1	-47.2	9.4
BIC	41.7	13.3	-26.6	81.9	71.9	-43.7	5.4

Note:

One scale, a 100-point scale, does not appear since there was only one effect size on that scale

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix B
Moderators

All moderator analyses, with the exception of political concordance, only reveal statistical associations, not causal effects, because the moderator variables vary mostly between studies: For example, some studies provided news sources, while others did not. But these studies differ in many other ways, all of which potentially confound any observed association.

Table B1 shows the results of the different meta regressions by moderator variable on discernment and Table B2 on response bias. Figures B1 and B2 visualize those results by showing the distribution of effect sizes by moderator variable.

Not preregistered moderators.

Scale symmetry. First, to avoid biasing our estimate for H2, we removed one study[56] that used a very asymmetrical set of answer options asked participants ("According to your knowledge, how do you rate the following headline? 1—not credible; 2—somehow credible; 3—quite credible; 4—credible; 5—very credible"). Second, we coded whether the remaining scales were perfectly symmetrical or not. Table B3 shows the frequency by which both scale types occurred.

Perfectly symmetrical scales include all binary scales (e.g. "True" or "False", "Real" or "Fake", is accurate "Yes" or "No", is accurate and unbiased "Yes" or "No"), and most Likert-scales (1 to 7: "Definitely fake" [1] to "Definitely real" [7], "Very unreliable" [1] to "Very reliable" [7], "Extremely unlikely" [1] to "Extremely likely" [7], "Extremely unbelievable" [1]

Table B1
Moderator effects on Discernment

	Country	Concordance	Family	Format	Source	Scale	Symmetrie	All
intercept	0.826***	0.584***	1.024***	0.957***	1.125***	1.112***	1.168***	0.013
	(0.080)	(0.103)	(0.082)	(0.085)	(0.106)	(0.089)	(0.087)	(0.039)
Country: US (vs. nonUS)	0.348**							1.125***
	(0.106)							(0.045)
Political Concordance : Discordant (vs. Concordant)		0.025						-0.004
		(0.049)						(0.046)
News family: Other (vs. Covid)			0.100					
			(0.156)					
News family: Political (vs. Covid)			-0.078					
			(0.111)					
News Format: Headline & Picture (vs. Headline)				0.007				-0.178
				(0.137)				(0.105)
News Format: Headline, Picture & Lede (vs. Headline)				0.097				
				(0.104)				
News source: Source (vs. No source)					-0.141			0.035
					(0.132)			(0.041)
Accuracy Scale: 6 (vs. 4)						-0.405**		
						(0.149)		
Accuracy Scale: 7 (vs. 4)						-0.028		
						(0.170)		
Accuracy Scale: binary (vs. 4)						-0.116		
						(0.128)		
Accuracy Scale: other (vs. 4)						-0.111		
						(0.134)		
Symmetrie: perfect (vs. imperfect)							-0.382***	-0.412**
							(0.106)	(0.096)
Num.Obs.	171	32	155	160	144	168	157	26
AIC	243.5	18.6	239.6	245.8	214.8	254.5	196.8	-19.3
BIC	256.0	24.4	254.8	261.2	226.6	276.4	209.0	-9.2

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

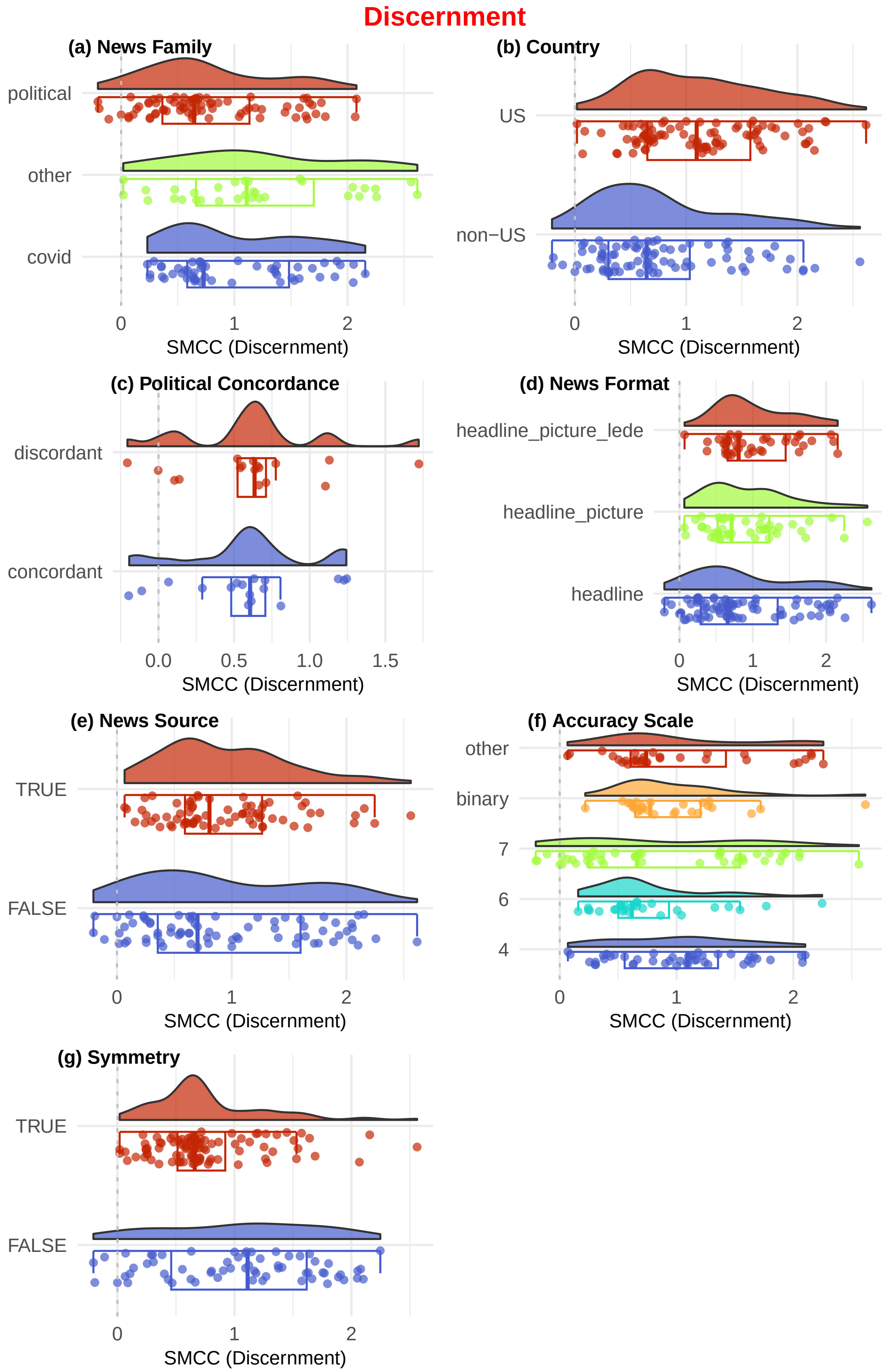

Figure B1. Distribution of effect sizes for discernment by moderator variables.

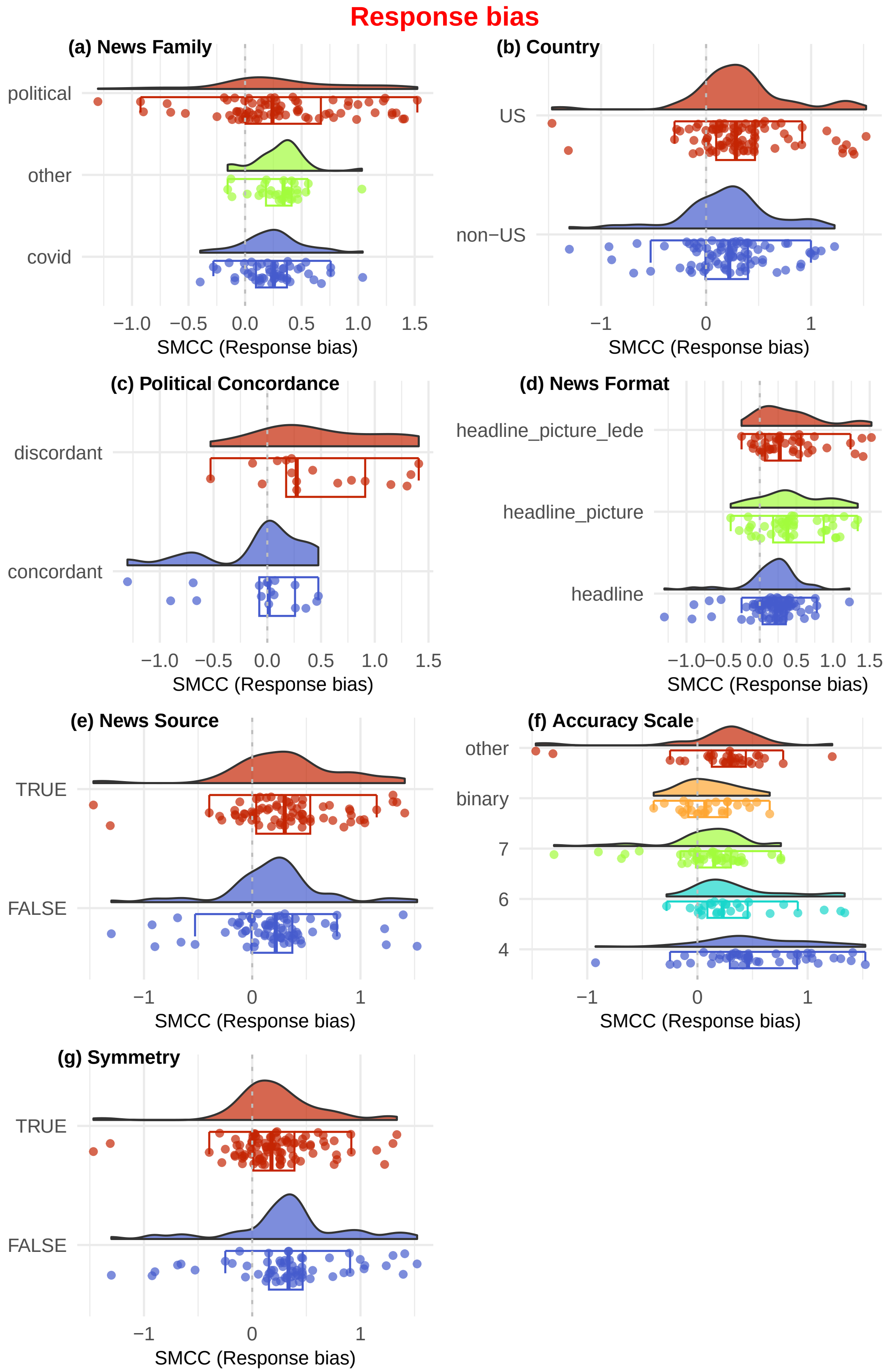

Figure B2. Distribution of effect sizes for response bias by moderator variables.

Table B2

Moderator effects on Response bias

	Country	Concordance	Family	Format	Source	Scale	Symmetrie	All
intercept	0.213***	-0.096	0.240***	0.179***	0.220***	0.542***	0.354***	-0.757**
	(0.054)	(0.124)	(0.043)	(0.047)	(0.062)	(0.077)	(0.071)	(0.190)
Country: US (vs. nonUS)	0.124+							1.067***
	(0.075)							(0.207)
Political Concordance : Discordant (vs. Concordant)		0.596***						0.639***
		(0.098)						(0.111)
News family: Other (vs. Covid)			0.079					
			(0.060)					
News family: Political (vs. Covid)			0.088					
			(0.082)					
News Format: Headline & Picture (vs. Headline)				0.254**				0.516**
				(0.082)				(0.125)
News Format: Headline, Picture & Lede (vs. Headline)				0.235**				
				(0.088)				
News source: Source (vs. No source)					0.096			0.306**
					(0.088)			(0.085)
Accuracy Scale: 6 (vs. 4)						-0.197+		
						(0.105)		
Accuracy Scale: 7 (vs. 4)						-0.460***		
						(0.104)		
Accuracy Scale: binary (vs. 4)						-0.403***		
						(0.093)		
Accuracy Scale: other (vs. 4)						-0.342**		
						(0.122)		
Symmetrie: perfect (vs. imperfect)							-0.142+	-0.747***
							(0.084)	(0.091)
Num.Obs.	171	32	155	160	144	168	157	26
AIC	218.7	46.7	185.1	178.9	201.3	201.9	206.3	22.0
BIC	231.3	52.6	200.3	194.3	213.2	223.8	218.5	32.0

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Table B3

	Imperfect Symmetry	Perfect symmetry	NA
Papers	19	25	4
Samples	50	68	13
Effects	64	95	14

Note. Frequency table of scales.

to “Extremely believable” [7]; 1 to 6: “Extremely inaccurate” [1] to “Extremely accurate” [6], “Completely false” [1] to “Completely true” [6]). Yet, we coded the most common scale, a 4-point Likert scale ([1] not at all accurate, [2] not very accurate, [3] somewhat accurate, [4] very accurate), as not perfectly symmetrical. We coded two other Likert scales as not perfectly symmetrical (“not at all trustworthy” [1] to “very trustworthy” [10]; “not at all” [1] to “very” [7]).

Third, we investigated whether H1 and H2 hold for both perfectly symmetrical and imperfectly symmetrical scales. While both H1 and H2 hold for both symmetry types, we found that studies with perfectly symmetric scales tend to yield lower discernment scores (Δ Discernment = -0.38 [-0.59, -0.17]) than studies relying on scales that are at least slightly asymmetric (Baseline discernment slightly asymmetric scales = 1.17 [1, 1.34]). We do not find a difference regarding response bias.

The results suggest that imperfectly symmetrical scales may inflate discernment.

However, the symmetry of response scales was not a factor that was experimentally manipulated, and the studies we compare in our model differ in many other ways and the observed difference is likely confounded.

Proportion of true news. Most studies exposed participants to 50% of false news and 50% of true news, whereas outside of experimental settings, people on average are exposed to much more true news than false news[52]. This inflated proportion of false news may increase discernment or make participants more skeptical of true news. Empirical evidence suggests that the ratio of false news has no effect on discernment and slightly increases skepticism in news judgment[5]. Figure B3 shows effect sizes for discernment and response bias as a function of news ratio. Due to the very uneven number of effect sizes, it does not seem reasonable to run a meta-regression to test this. However, Fig. B3 suggests no obvious trend with regard to the share of true news ratio. Besides, as for the other moderator variables, any observed association is likely to be confounded by other factors.

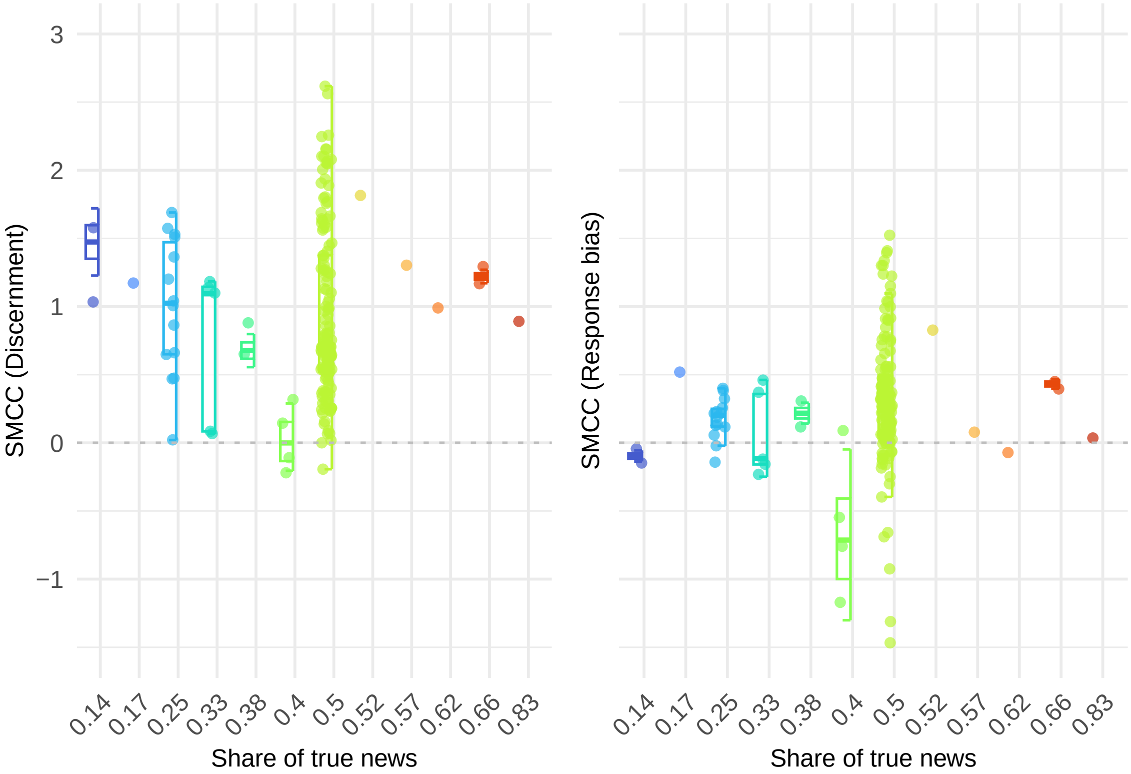

Figure B3. Effect sizes plotted by their share of true news among all news that an individual participant saw.

Appendix C
Individual level data

We compare the results of our main meta model to the individual-level data with the following procedure: First, we restrict our data to (i) only studies using a non-binary response scale and (ii) only those studies that we have individual-level data on. Second, we run the same meta-analytic model as in the main analysis on the effect sizes of that subset of studies. Third, we take the individual-level data of that subset of studies and run a mixed model on it.

The meta-model estimates are standardized. To be able to compare results, we standardized participants' accuracy ratings in the individual-level data as follows: Within each sample, we calculated the standard deviation of accuracy ratings (false and true news combined). Then, for each sample, we divided accuracy ratings by the respective standard deviation.

We use the `lme4` package[69] and its `lmer()` function to run the mixed models. The mixed models include random effects by participant (each participant provides several ratings for both true and false news) and by sample for both the intercept and the effect of veracity. In our models, participants are nested in samples.

As shown in Fig. C1, this individual-level analysis yields an estimate very similar to our meta-analytic average.

How skilled were individual participants?

In our meta analysis, we find that people discern well between true and false news - on average. But how skilled are individuals in discerning true from false?

We calculate a discernment score per individual for all individual level studies. To compare across different scales, we transpose all accuracy scores on a scale from 0 to 1, resulting in a range of possible values from -1 to 1 for both discernment and response bias (see C2).

We report the absolute number of individuals with a positive vs. negative discernment and response bias score in Table C1.

Table C1

	Discernment	Response bias
negative	1586 (0.125)	4573 (0.362)
positive	11062 (0.875)	8075 (0.638)

Note. Frequency table of total number of participants that had a positive or negative score for both outcomes.

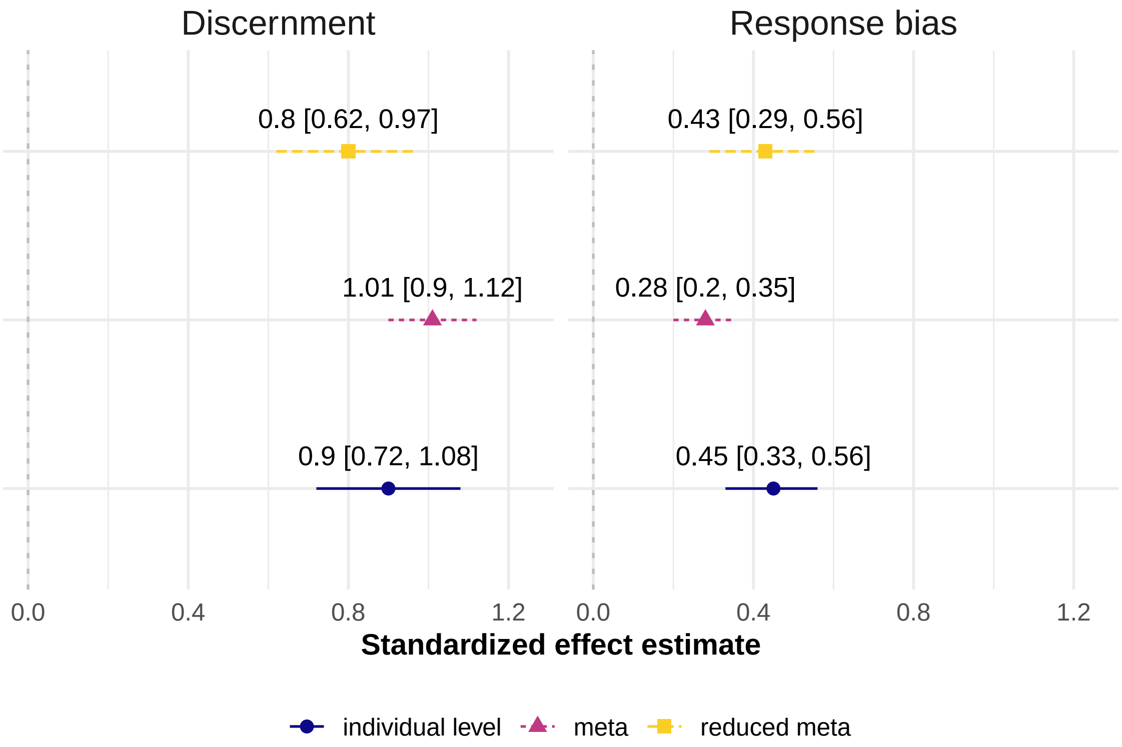

Figure C1. Comparison of meta to individual level analysis (continuous scales only). "Meta" corresponds to the main results reported in the paper; "meta reduced" are the same meta-analytic models as in the main analysis but run on the subset of studies for which we have individual level data; "individual-level" corresponds to the result of mixed effect models run on the individual-level data. Symbols represent estimates, horizontal bars 95% confidence intervals.

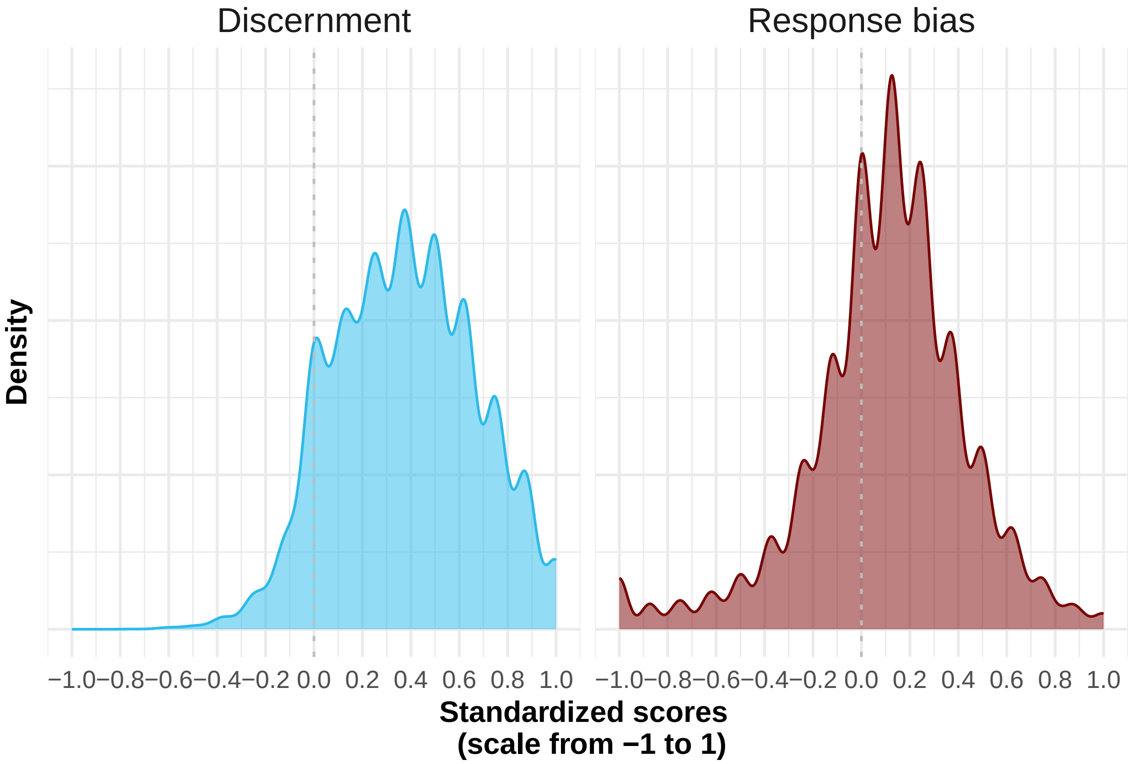

Figure C2. Distribution of average discernment and response bias scores per individual participant in the subset of studies that we have raw data on. We standardized original accuracy ratings to range from 0 to 1, to be able to compare across scales. Therefore, the worst possible score is -1 where, for discernment, an individual classified all news wrongly, and for response bias, an individual classified all true news correctly (as true) and all false news incorrectly (as true). The best possible score is 1 where, for discernment, an individual classified all news correctly, and for response bias, an individual classified all true news incorrectly (as false) and all false news correctly (as false).

Appendix D
Binary vs. continuous scales

Some of the studies included in our review measure perceived accuracy on a continuous scale, others on a binary (or dichotomous) scale. This is not problematic per se - there are statistical methods to compare effects on both scales[65]. These require, however, appropriate summary statistics for both scales. For continuous measures, means and standard deviations are fine; for binary measures we would need, for example, odds or risk ratios. The problem we were facing is that authors did not provide the appropriate summary statistics for binary scales. Instead, they tended to report means and standard deviations, just as they do for continuous outcomes. For the main analysis, we made the decision to treat continuous and binary scales in the same way, glossing over potential biases from inappropriate summary statistics. Here, we include robustness checks to see how this decision affects our results. First, we ran meta-regression to see if there are any differences regarding our outcome variables associated with type of scale (continuous vs. binary). This analysis suggests that there is a difference regarding response bias between studies using binary and continuous scales (smaller bias for binary scales). Second, we focus on the subset of studies that we have individual-level data on. For this subset, we combine binary and continuous response scales by collapsing the latter on a binary outcome. We calculate appropriate effect sizes for binary data, namely log odds ratios (logORs) and run a meta-analysis on those effect sizes. The results are in line with our main findings, suggesting positive discernment and response bias.

Meta-regression

We ran a meta-regression using scale type (two levels: binary vs. continuous) as a predictor variable. Table D1 summarizes the results, and Fig. D1 illustrates them. The analysis suggests that response bias is more enhanced among continuous studies. However, we cannot tell how much of that observed difference is due to relying on imperfect summary statistics for binary scale, or due to other factors.

Table D1
Model results

	Discernment	Response bias
intercept	1.001***	0.147**
	(0.079)	(0.052)
Continuous (vs. binary)	0.013	0.154*
	(0.076)	(0.070)
Num.Obs.	171	171
AIC	254.1	219.2
BIC	266.6	231.7

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

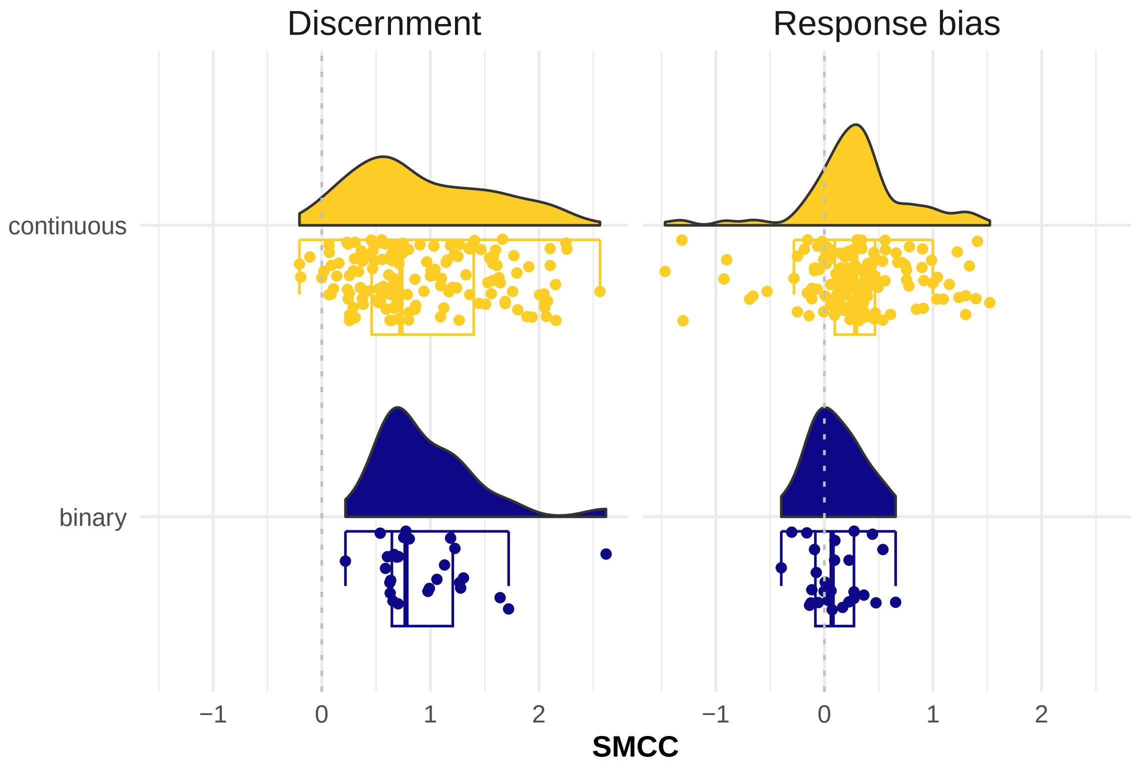

Figure D1. Distribution of effect sizes (SMCC) grouped by whether a binary or continuous response scale was used.

Individual-level data

Here, we focus on the subset of studies that we have raw, individual-level data on. For this subset of studies, we first calculated the odds ratios from the raw data[6]. We then ran a meta-analysis on the odds ratios.

Odds ratios. The 'odds' refer to the ratio of the probability that a particular event will occur to the probability that it will not occur, and can be any number between zero and infinity[65]. It is commonly expressed as a ratio of two integers. For example, in a clinical context, 1 out of 100 patients might die; then the odds of dying are `0.01`, or `1:100`.

The odds *ratio* (OR) is the ratio of the Odds. The odds ratio that characterizes discernment is calculated as

$$OR_{Accuracy} = \frac{(Accurate_{true}/NotAccurate_{true})}{(Accurate_{false}/NotAccurate_{false})}$$

If the OR is `1`, participants were just as likely to rate items as 'accurate' when looking at true news as they were when looking at false news. If the OR is `> 1`, then participants

[6] A general overview of appropriate summary statistics for binary outcomes can be found here([65]): https://training.cochrane.org/handbook/current/chapter-06#section-6-4)

Table D2

Veracity	Rated as accurate	Rated as not accurate	Sum
fake	10703 (0.251)	32001 (0.749)	42704 (1)
true	29771 (0.698)	12872 (0.302)	42643 (1)

Note. Frequency of responses (among individual-level studies with binary response scales)

rated true news as more accurate than fake news. An OR of 2 means that participants were twice as likely to rate true news as accurate compared to false news.

The OR for response bias is calculated as

$$OR_{Error} = \frac{(NotAccurate_{true}/Accurate_{true})}{(Accurate_{false}/NotAccurate_{false})} = \frac{\frac{1}{(NotAccurate_{true}/Accurate_{true})}}{(Accurate_{false}/NotAccurate_{false})} = \frac{1}{OR_{Accuracy}}$$

For our analysis, we calculated the odds ratio (OR) for both accuracy and error. More precisely, we expressed the OR on a logarithmic scale, also referred to as “log odds ratio”(logOR). As for odds ratios, if the log odds ratio is positive, it indicates positive discernment/response bias[7].

Table D2 shows the frequency of answers by veracity.

Meta-analysis. We ran a meta-regression on the odds ratios. The results can be found in Table D3. For reference, we also present the results for that subset using the main analysis estimator (SMCC) and a non-standardized estimator that likewise accounts for dependence between false and true news, namely the mean change (MC)[8]. The results suggest that when using odds ratios, we do not find a statistically significant response bias. However, this analysis relies on very few observations (6 effect sizes only), hence low statistical power to detect a potential effect. We therefore extended the analysis by adding also individual-level studies with continuous response scales and collapsing ratings to a binary outcome. For example, on a 4-point scale, we coded responses of 1 and 2 as not accurate (0) and 3 and 4 as accurate (1). For scales, with a mid-point (example 3 on a 5-point scale), we coded midpoint answers as NA. The results of this extended analysis can be found in table D4. In line with our main results, this extended analysis suggests a positive response bias.

[7] To interpret the magnitude of that difference we have to transform the logarithmic estimate back to a normal odds ratio. The reason we use the log odds ratios in the first place is that which makes outcome measures symmetric around 0 and results in corresponding sampling distributions that are closer to normality[63]

[8] We use the term mean change in line with vocabulary used by the metafor package and its `escalc()` function that we use for all effect size calculations. It is in fact a simple mean difference but one that accounts for the correlation between true and false news in the calculation of the standard error (see[65]). Here is a direct link to the relevant chapter online: https://training.cochrane.org/handbook/current/chapter-23#section-23-2-7-1

Table D3

Individual-level studies with binary response scale

	(based on individual data)		*(based on meta data)*			
	Outcome: Log OR		Outcome: SMCC		Outcome: MC	
	Accuracy	Error	Accuracy	Error	Accuracy	Error
Estimate	1.850***	0.197+	0.759***	0.156*	0.432***	0.088*
	(0.175)	(0.110)	(0.071)	(0.076)	(0.030)	(0.043)
Num.Obs.	6	6	12	12	12	12
AIC	14.1	7.7	-20.9	0.8	-39.6	-14.1
BIC	13.4	7.1	-19.5	2.2	-38.2	-12.6

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: Note that the number of observations differ, because some samples provide several effect sizes in the meta-data. For the odds ratios based on the individual data, however, we calculated only one average effect size per sample. The samples are only from studies with binary response scales that we had raw, individual-level data on.

Table D4

Individual-level studies with likert scale ratings collapsed to binary outcome

	(based on individual data)		*(based on meta data)*			
	Outcome: Log OR		Outcome: SMCC		Outcome: MC	
	Accuracy	Error	Accuracy	Error	Accuracy	Error
Estimate	1.985***	0.494***	0.789***	0.340***	0.971***	0.517***
	(0.129)	(0.095)	(0.062)	(0.056)	(0.093)	(0.110)
Num.Obs.	21	21	38	38	38	38
AIC	45.4	32.1	9.1	33.2	44.6	83.2
BIC	48.5	35.2	14.0	38.1	49.5	88.2

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: Note that the number of observations differ, because some samples provide several effect sizes in the meta-data. For the odds ratios based on the individual data, however, we calculated only one average effect size per sample. The sample consists of all studies we had individual-level data on. For individual-level studies with continuous response scales, we computed the odds ratio after collapsing responses to a binary outcome.

Appendix E
Publication bias

To quantify asymmetry as visualized by the funnel plot, we ran Egger's regression test[70] following our pre-registration. The results are displayed in Table E1. The outcome variable in the Egger's regression test is the observed effect size divided by its standard error. The resulting value is a z-score, which tells us directly if an effect size is significant: If $z \geq 1.96$ or $z \leq -1.96$, we know that the effect is significant ($p < 0.05$). This outcome is regressed on the inverse of its standard error, a measure of precision, with higher values indicating higher precision[66]. The coefficient of interest in the Egger's test is the intercept, i.e. the estimated z-score when precision (the predictor variable) is zero. Given a precision of 0, or an infinitely large standard error, we would expect a z-score scattered around 0. However, when the funnel plot is asymmetric, for example due to publication bias, we expect that small studies with very high effect sizes will be considerably over-represented in our data, leading to a surprisingly high number of low-precision studies with high z-values. Due to this distortion, the predicted value of y for zero precision will be considerably larger than zero, resulting in a significant intercept. However, just as asymmetries in the funnel plot can stem from sources of heterogeneity other than publication bias, a positive Egger's regression is not proof for publication bias. In fact, because we had no a priori suspicion of publication bias - our outcomes have not been of the outcomes of interest in the original studies - we do not take the results of the Egger's test as indicative of publication bias.

Table E1
Egger's regression

	Discernment	Response bias
(Intercept)	24.093***	1.284
	(6.615)	(5.830)
Inverse SE	0.528***	0.270***
	(0.080)	(0.062)
Num.Obs.	173	173
R2	0.202	0.099
R2 Adj.	0.197	0.094
AIC	1776.2	1721.6
BIC	1785.7	1731.0
Log.Lik.	-885.115	-857.785
RMSE	40.34	34.44

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Appendix F
Country comparison

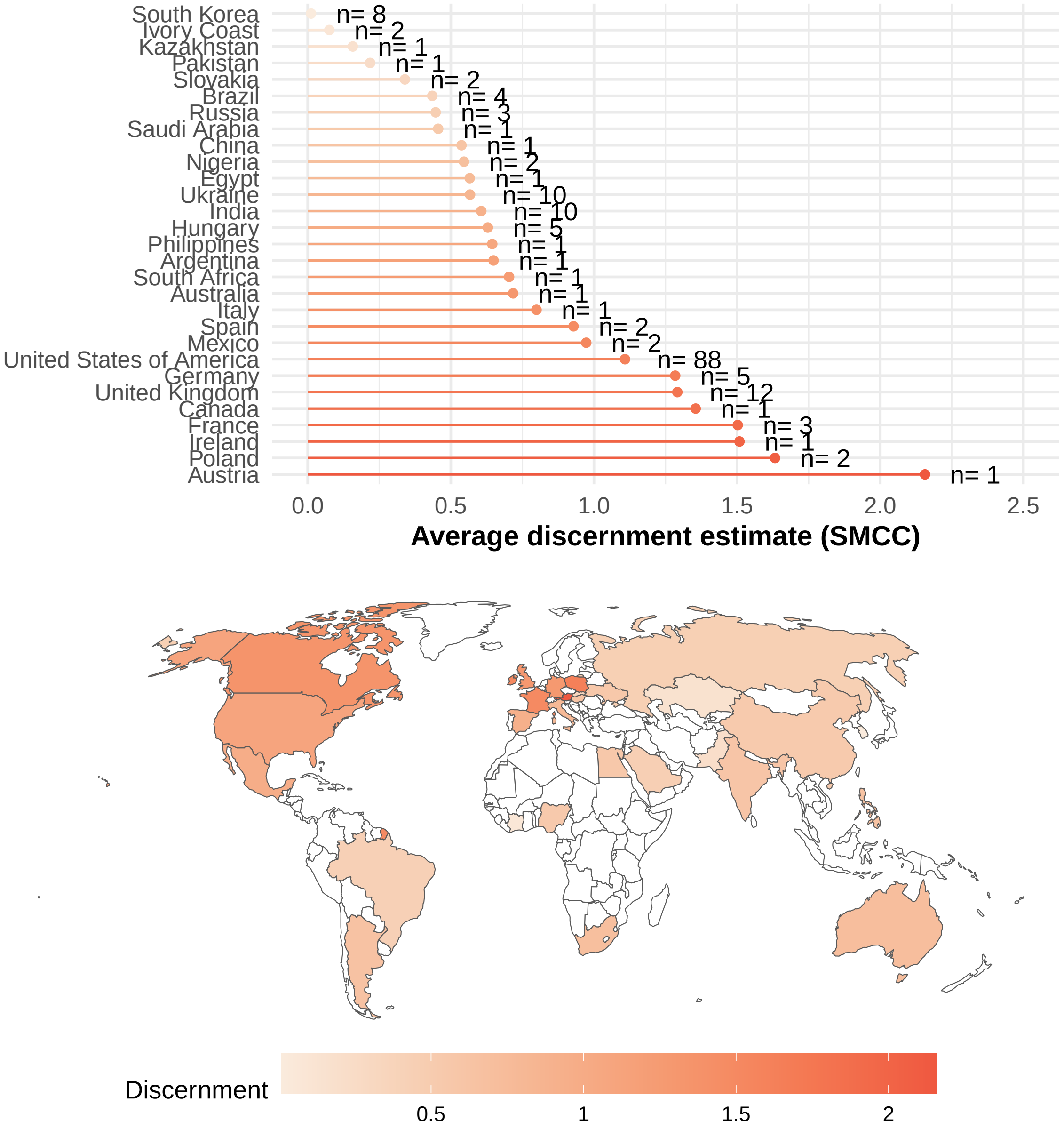

Figure F1. Discernment estimates by country.

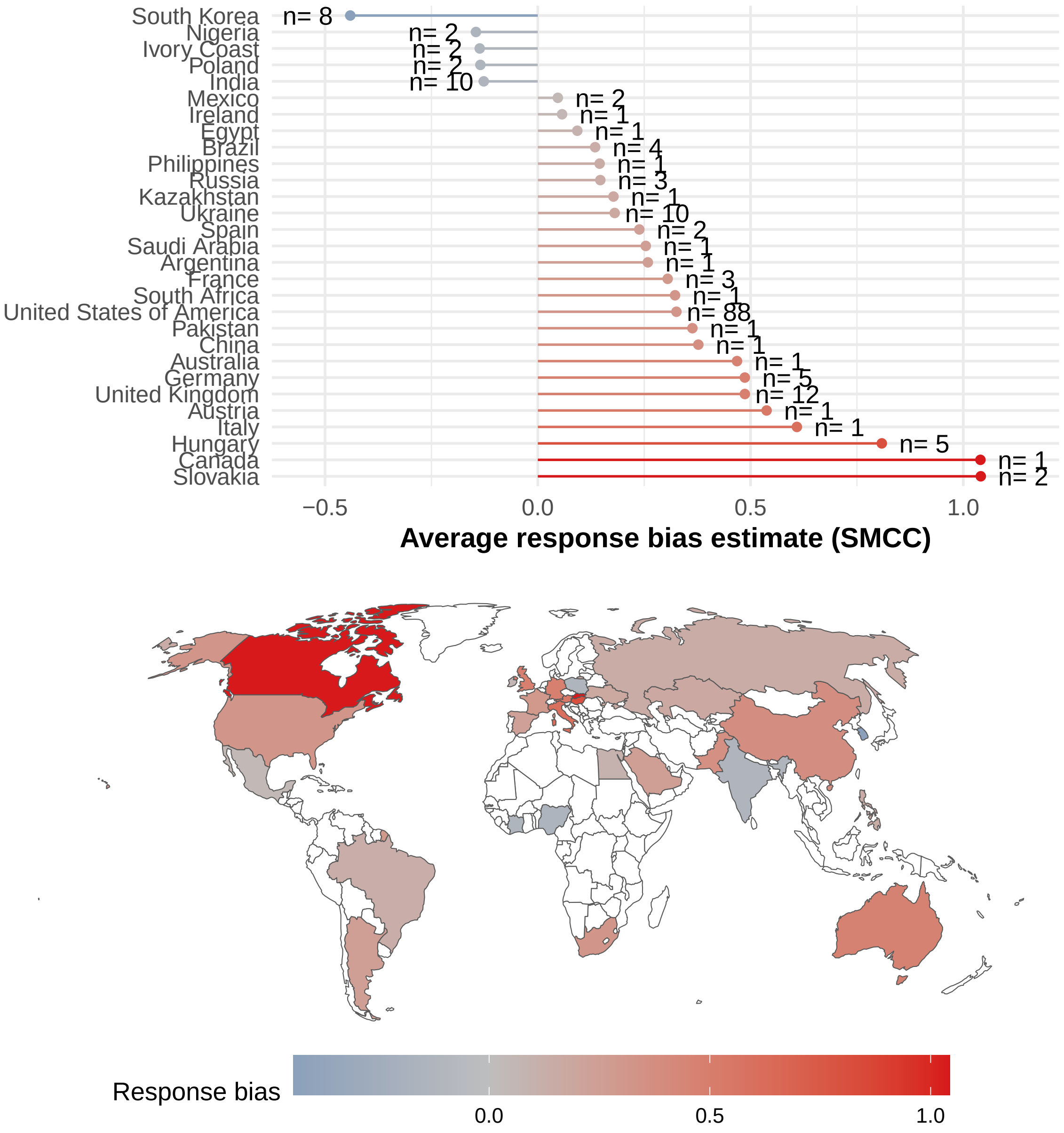

Figure F2. Response bias estimates by country.

Appendix G
Selection bias

Response bias could be an artifact of biased news selection for experiments. For example, one might suspect researchers to pick easy-to-detect false news and/or hard-to-detect true news (e.g. to avoid ceiling effects), thus inflating participants skepticism of true news.

We deem this unlikely because of several reasons: First, while ceiling effects may have biased the selection of 'non-obviously true' true news, floor effects should have conversely biased the selection of 'non-obviously false' false news. We see no reason to suspect that news selection has created accuracy asymmetries between false and true news items. Researchers did sometimes pre-test items to avoid using excessively ambiguous headlines. Yet, no article mentions asymmetries in this relatively rare selection process.

Second, degrees of accuracy could be captured by Likert scales, but not by binary scales. If a selection bias is mild, i.e. leading to nuances within veracity categories but not to misclassification (e.g. people rating a true item as 5 instead of 7 on a 7-point accuracy scale, but not as a 3), then collapsing responses from Likert scales to binary outcomes should account for it. In Appendix D, we show that our results hold when doing so for the subset of studies for which we have individual-level data.

Third, and most importantly, we observe similar average accuracy ratings for true news in studies that randomly sampled true news from high-quality mainstream news sites. These samples of headlines are free of any selection bias that may originate from researchers selecting not obviously accurate true headlines.[71] used CrowdTangle to automatically scrap 500 headlines from 20 mainstream news sites and had participants rate the accuracy of these headlines. The mean accuracy rating of these headlines was 5.05 (sd = 0.56) on a 7-point scale, or 0.68 if we transpose the scale to reach from 0 to 1. This is similar to our (unweighed) average true news rating (0.60) when scaling effect sizes to range from 0 to 1 (see Fig. 2). Similarly,[72] automatically scrapped true headlines using the Google News API. On a 7-point scale, the average true news rating was 4.45 (sd = 1.66), or 0.57 on a scale from 0 to 1.

Finally, we briefly discuss an important recent working paper that nuances some of our findings.[73] automatically scrapped popular headlines on reliable and unreliable websites in the US over a period of one month. They found that participants discerned between true and false/misleading headlines, but that they were better at rating true headlines as true than false/misleading headlines as false/misleading (suggesting a negative response bias). We believe that the discrepancies between their findings and the ones of our meta-analysis boil down to the sample of false news. The papers included in our meta-analysis almost exclusively rely on false news identified as such by fact-checking websites. By contrast,[73] hired a team of fact checkers to verify news items right after their publication. We suspect that this results in a sample of false news that are harder-to-detect as such than the ones in our samples for two reasons: First, since the news were not yet fact-checked, they might simply be less well known to be false. Second, there might be a selection bias, such that fact-checking websites verify blatantly false news first, because debunking them is easier. If this is the case, their false news items are harder to detect than the ones that our meta-analysis is based on.