A Signal Detection Approach to Understanding the Identification of Fake News

Cédric Batailler

Université Grenoble Alpes

Skylar M. Brannon

University of Texas at Austin

Paul E. Teas

University of Illinois at Chicago

Bertram Gawronski

University of Texas at Austin

Abstract

Research across many disciplines seeks to understand how misinformation spreads with a view towards limiting its impact. One important question in this research is how people determine whether a given piece of news is real or fake. The current article discusses the value of Signal Detection Theory (SDT) in disentangling two distinct aspects in the identification of fake news: (1) ability to accurately distinguish between real news and fake news and (2) response biases to judge news as real versus fake regardless of news veracity. The value of SDT for understanding the determinants of fake news beliefs is illustrated with reanalyses of existing data sets, providing more nuanced insights into how partisan bias, cognitive reflection, and prior exposure influence the identification of fake news. Implications of SDT for the use of source-related information in the identification of fake news, interventions to improve people's skills in detecting fake news, and the debunking of misinformation are discussed.

Keywords: cognitive reflection, illusory truth effect, misinformation, partisan bias, signal detection theory

Misinformation comes in various forms ranging from the more entertaining, such as satirical pieces from *The Onion*, to the more insidious, such as Nazi propaganda and fabricated reports suggesting a link between vaccinations and autism. Although fake news is not a new concept, concerns over the impact of misinformation have grown considerably as the internet and social media provide a conduit for spreading information widely and rapidly, regardless of its veracity. Because false information often continues to impact judgments and decisions even after being refuted (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Rapp & Braasch, 2014; Schwarz, Sanna, Skurnik, & Yoon, 2007; for a meta-analysis, see Chan, Jones, Jamieson, & Albarracín, 2017), exposure to misinformation poses a major challenge for the functioning of societies in the so-called information age. Given the growing concerns over the dangers of misinformation (Mitchell, Gottfried, Stocking, Walker, & Fedeli, 2019), researchers across many disciplines are trying to understand how misinformation spreads with a view towards limiting its impact (Lazer et al., 2018). For example, research in the computer sciences has focused on building algorithms that predict, flag, and block sources of misinformation online (see Conroy, Rubin, & Chen, 2015). Research in the social sciences, on its part, has focused on understanding what factors contribute to belief in misinformation and effective routes to reducing its impact (see Lewandowsky et al., 2012).

Although research in psychology has made significant progress in understanding the factors that influence people's belief in misinformation (for reviews, see Lewandowsky et al., 2012; Rapp & Braasch, 2014), studies on how people determine the veracity of news have relied on approaches that conflate two conceptually distinct aspects in the identification of fake news: (1) ability to accurately distinguish between real news and fake news and (2) response biases to judge news as real versus fake regardless of news veracity. In the current article, we discuss how

Signal Detection Theory (SDT; Green & Swets, 1966) can provide more nuanced insights into the processes underlying the propagation of fake news by disentangling the two aspects. We illustrate the value of SDT with reanalyses of existing data sets, uncovering the particular manner in which various factors influence the identification of fake news.

Identifying Fake News

A fundamental question in research on the effects of fake news is how people determine whether a piece of information is real or fake. Guided by different theoretical frameworks, prior research on this question has focused on four determinants: (1) partisan bias, (2) cognitive reflection, (3) motivated reflection, and (4) prior exposure.

Partisan Bias

One important factor in the identification of fake news is the congruence versus incongruence of (mis)information with prior beliefs. According to motivational accounts that emphasize the significance of ideological beliefs for social identities, people have a tendency to accept information that is congruent with their ideological beliefs and dismiss information that is incongruent with their ideological beliefs (e.g., Van Bavel & Pereira, 2018). Importantly, acceptance of ideology-congruent information and rejection of ideology-incongruent information is assumed to occur independently of the actual veracity of the relevant information, leading people to accept fake news that is congruent with their ideological beliefs and dismiss real news that is incongruent with their ideological beliefs. For example, supporters of a particular politician may accept fake news that sheds a positive light on that politician and dismiss real news as fake news if it sheds a negative light on that politician. Conversely, critics of the same politician may accept fake news that sheds a negative light on that politician and dismiss real news as fake news if it sheds a positive light on that politician.

A similar prediction is implied by cognitive accounts suggesting that people use consistency as a cue to judge the validity of information (see Gawronski, 2012; Schwarz & Jalbert, 2020). These accounts similarly suggest that people have a tendency to judge new information as valid if it is consistent with prior beliefs. Moreover, when new information is inconsistent with prior beliefs, people often reconcile the inconsistency by generating an explanation for the new information that reconciles its inconsistency with prior beliefs (Johnson-Laird, Girotto, & Legrenzi, 2004). Because dismissing ideology-incongruent news as fake is an effective strategy to resolve its inconsistency with prior ideological beliefs, cognitive consistency accounts similarly suggest that people tend to accept fake news that is congruent with their ideological beliefs and dismiss real news as fake news if it is incongruent with their ideological beliefs.

Cognitive Reflection

In contrast to accounts emphasizing the impact of prior ideological beliefs, other accounts suggest that people's susceptibility to fake news is driven by belief-unrelated differences in cognitive reflection. According to these accounts, belief in fake news reflects insufficient analytic thinking rather than partisan bias. In line with this hypothesis, some research suggests that people's ability to correctly identify fake news is associated with individual differences in cognitive reflection, in that individuals with higher scores on the Cognitive Reflection Test (CRT; Frederick, 2005) were more accurate in distinguishing between real news and fake news than individuals with lower scores on the CRT (Pennycook & Rand, 2019). Importantly, this relation held regardless of the political slant of the news, in that higher CRT scores were associated with greater accuracy regardless of whether the news was congruent or incongruent with participants' political leaning. Similar results were obtained in studies using experimental manipulations of reflective thinking (Bago, Rand, & Pennycook, 2020). Thus, applied to the above example, any factor that supports cognitive reflection should increase a person's accuracy in identifying fake news about a particular politician regardless of whether the person supports or opposes that politician.

Motivated Reflection

In contrast to accounts that treat cognitive reflection and partisan bias as mutually exclusive factors, other accounts suggest that the two factors can interactively determine belief in fake news. Based on the idea that people employ cognitive processes in the service of their goals (Ditto & Lopez, 1992; Kruglanski & Webster, 1996; Kunda, 1990), motivated-reflection accounts suggest that people strategically utilize their cognitive skills to process information in a manner such that the inferential outcomes are consistent with beliefs they are motivated to protect (Kahan, Peters, Dawson, & Slovic, 2017). According to this view, people are often motivated to reach conclusions that support their ideological beliefs, and success in accomplishing this inferential goal depends on basic cognitive skills (e.g., intelligence, literacy, numeracy). In such cases, partisan bias in the identification of fake news should increase (rather than decrease) as a function of basic cognitive skills (Kahan, 2017). That is, people with greater reflective abilities should show a stronger tendency to accept ideology-congruent information and dismiss ideology-incongruent information compared to people with weaker reflective abilities. Thus, applied to our thematic example, supporters of a particular politician may accept fake news that sheds a positive light on that politician and dismiss real news as fake news if it sheds a negative light on that politician, and this partisan bias should be more pronounced among people with stronger reflective abilities.

Prior Exposure

Another important factor in judgments of veracity is processing fluency. A considerable body of research suggests that people use the experienced fluency of processing information as a meta-cognitive cue for judging the veracity of that information, in that people treat high fluency as an indicator of accuracy (Reber & Unkelbach, 2010). An important determinant of fluency is prior exposure, which has been found to increase perceptions of veracity by increasing the ease of processing the relevant information (Lewandowsky et al., 2012; Schwarz et al., 2007; Unkelbach, Koch, Silva, & Garcia-Marques, 2019). Applied to the current question, fluency accounts suggest that prior exposure to fake news increases the ease of processing its content, which increases perceptions of veracity (Schwarz & Jalbert, 2020). In line with this idea, Pennycook, Cannon, and Rand (2018) found that prior exposure to fake news headlines increased the likelihood that the headlines were judged as real, and this effect was unaffected by the congruence of the headlines with participants' political ideology. These findings resonate with the claims of purely cognitive accounts, suggesting that belief in fake news is rooted in basic cognitive processes rather than motivated reasoning. Thus, applied to our thematic example, prior exposure to a fake news article about a particular politician may increase the likelihood that people perceive the news article as real regardless of whether the article's content is congruent or incongruent with the reader's political leaning.

Signal Detection Theory

Although previous research has provided valuable insights into the factors that influence people's acceptance of misinformation, many studies in this area have conflated two conceptually distinct aspects in the identification of fake news: (1) ability to accurately distinguish between real news and fake news and (2) response biases to judge news as real versus fake regardless of news veracity. Because discrimination accuracy and responses biases are likely rooted in different underlying processes, conflating the two aspects can lead to incorrect conclusions about the psychological determinants of fake news beliefs. SDT offers a simple and effective way to disentangle discrimination accuracy and response bias by providing independent indices for the two aspects. In this section, we briefly review the core ideas underlying SDT and discuss how its application to the identification of fake news can provide more nuanced insights into the determinants of fake news beliefs.

The use of SDT originated in perceptual studies to understand how different factors influence people's ability to distinguish signals from noise (Green & Swets, 1966). Since then, SDT has been applied to a wide range of topics in psychology, including recognition memory and racial bias in weapon identification. A common feature of these applications is that they are concerned with the same basic question: How well can people distinguish between two classes of stimuli? For example, in studies on recognition memory, how well can people distinguish words that have been presented in a prior task from words that have not been presented before (Snodgrass & Corwin, 1988)? In studies on racial bias in weapon identification, how well can people distinguish weapons from non-threatening objects (Payne & Correll, 2020)? Applied to fake news, how well can people discern fake news from real news?

One possible approach to answer these questions is to focus on hits: cases in which participants correctly identify the focal target stimuli (e.g., correct classification of previously presented words, weapons, or fake news articles). However, simply tallying a participant's hits ignores that two independent mechanisms can lead to correct classifications of target stimuli. First, participants may correctly classify the target stimuli because they are able to accurately distinguish the signal from the noise. For example, in studies on recognition memory, participants may correctly identify previously presented words because they are able to accurately distinguish previously presented words from new lures; in studies on racial bias in weapon identification, participants may correctly identify weapons because they are able to accurately distinguish weapons from non-threatening objects; and in studies on the identification of fake news, participants may correctly identify fake news articles because they are able to accurately distinguish fake news from real news. Second, participants may correctly classify the target stimuli because they have a tendency to respond *yes, this stimulus fits the focal parameters* regardless of whether the stimulus actually fits those parameters. For example, in studies on recognition memory, participants may respond *old* for all words regardless of whether they were presented before; in studies on racial bias in weapon identification, participants may respond *weapon* for both weapons and non-threatening objects; and in studies on the identification of fake news, participants may respond *fake* for all news articles regardless of their veracity.

Although both of these factors lead to a “hit” in identifying the presence of a target stimulus, they represent fundamentally distinct patterns of responses with distinct underlying mechanisms. Thus, confounding them in overall hit rates can lead to inaccurate interpretations of the data. SDT offers a simple means to disentangle the two aspects by providing separate indices for each aspect as a function of an individual’s hits (e.g., correct classification of previously presented words, weapons, or fake news articles) and false alarms (e.g., incorrect classification of new foil words as having been presented before, non-threatening objects as weapons, or real news articles as fake).

SDT's index for discrimination sensitivity (labeled *d′*) reflects the distance between the distributions of judgments about two stimulus classes along the judgment-relevant dimension.[1] For example, when judging news articles as real (vs. fake), *d′* indicates the difference in the distributions for real news versus fake news along the dimension of perceived veracity (see Figure 1).[2] Distributions that are further apart along the perceived veracity dimension have a higher *d′*, indicating that participants' ability in correctly discriminating between real news and fake news is relatively high. Conversely, distributions that are closer together along the perceived veracity dimension have a lower *d′*, indicating that participants' ability in correctly discriminating between real news and fake news is relatively low. Indeed, if the distributions for real news and fake news overlap on the perceived veracity dimension, some real news might be perceived as "less real" than fake news and some fake news might be perceived as "more real" than real news (see Figure 1). Conceptually, factors that decrease *d′* pull the distributions closer together, making it more difficult to discriminate stimuli from each class. Conversely, factors that increase *d′* pull the distributions further apart, making it easier to discriminate stimuli from each class. Mathematically, discrimination sensitivity is captured by the difference between a participant's hit rate and false-alarm rate:

$$d' = \mathrm{z(H)} - \mathrm{z(FA)}$$

In this equation, *H* refers to hit rate or the proportion of target trials on which a participant showed the correct response (e.g., number of *real* classifications of real news articles

[1] In its most popular variant, SDT assumes that the distributions for targets and lures have the same variance (see Figure 1). Unequal variance can be accounted for in a variant of SDT that uses different indices to quantify discrimination sensitivity and response bias (Green & Swets, 1966).

[2] Although research on fake news detection has focused primarily on categorical differences between real news and fake news, it is worth noting that misinformation spread by news outlets can also come in variants that do not qualify as fake news (e.g., hyper-partisan news with misleading, but not entirely incorrect content).

divided by the total number of real news articles; see Table 1); *FA* refers to false-alarm rate or the proportion of distracter trials on which a participant showed the incorrect response (e.g., number of *real* classifications of fake news articles divided by the total number of fake news articles; see Table 1). Both H and FA follow a quantile function for a z distribution (or inverse cumulative distribution function) in a manner such that a proportion of 0.5 is converted to a z-score of 0 (reflecting chance responses). Thus, proportions greater than 0.5 (i.e., above-chance responses) produce positive z-scores and proportions smaller than 0.5 (i.e., below-chance responses) produce negative z-scores. Extreme *d′* scores occur when participants show near-perfect accuracy. For example, if H = .99 and FA = .01, *d′* = 4.65. For perfect accuracy (i.e., H = 1.00 and FA = 0.00), *d′* is infinite, requiring adjustments before the calculation of *d′* scores.[3]

SDT's index for response bias (labeled *c*) reflects the threshold along the judgment-relevant dimension at which a participant decides to switch their decision. For example, when judging whether news articles are real (vs. fake), *c* indicates the degree of veracity one must perceive before judging a news article as real (see Figure 2). Any stimulus with greater perceived veracity than that value will be judged as real, whereas any stimulus with lower perceived veracity than that value will be judged as fake. In this example, a higher (or more conservative) criterion would indicate that a participant is generally less likely to judge a news story as real, whereas a lower (or more liberal) criterion would indicate that a participant is generally more likely to judge a news story as real. Mathematically, response bias (or threshold) is captured by the following equation:

[3] For such cases, MacMillan and Creelman (2004) suggest to "convert proportions of 0 and 1 to 1/(2*N*) and 1 − 1/(2*N*), respectively, where *N* is the number of trials on which the proportion is based" (p. 8). An alternative strategy is to "add 0.5 to all data cells regardless of whether zeroes are present" (p. 8).

$$c = -1 \times \frac{\mathrm{z(H)} + \mathrm{z(FA)}}{2}$$

When the false-alarm rate is equal to the rate of misses (see Table 1), c equals 0, because z(FA) = z(1−H) = −z(H) (see Macmillan & Creeman, 2004). Negative c values arise when the false-alarm rate is greater than the miss rate, and positive values arise when the false-alarm rate is smaller than the miss rate (see Table 1). Extreme c values occur when H and FA are both large or both small. For example, if both H and FA are .99, $c = -2.33$. In contrast, if both H and FA are .01, $c= +2.33$.

Although d' and c are both based on hits and false alarms, the two indices are conceptually independent from one another (see Macmillan & Creelman, 2004; Stanislav & Todorov, 1999), which means that any given factor can influence either d' or c, or both. This aspect is important, because a closer examination of the reviewed factors in the identification of fake news reveals that they are not mutually exclusive. When analyzed from the perspective of SDT, partisan bias should be evident in response bias scores (c), in that people should show a lower threshold for judging news articles as real when they are congruent than when they are incongruent with their ideological beliefs. In contrast, the proposed effect of cognitive reflection should be evident in discrimination sensitivity scores (d'), in that greater cognitive reflection should be associated with a stronger ability to distinguish real news and fake news. Moreover, the proposed effect of motivated reflection should be evident in response bias scores (c), in that the tendency to show a lower veracity threshold for ideology-congruent than ideology-incongruent news should be more pronounced for people with stronger reflective abilities. Finally, prior exposure may influence judgments either by reducing people's ability to accurately discriminate between real news and fake news (d') or by increasing the tendency to judge news articles as real regardless of their veracity (c), or both.[4]

The Value of SDT for Studying the Identification of Fake News

To illustrate the insights SDT can provide for research on the identification of fake news, we reanalyzed data sets from two published articles on fake news discernment (Pennycook et al., 2018; Pennycook & Rand, 2019). In the first article, Pennycook and Rand (2019) investigated the role of cognitive and motivational factors in the identification of fake news. In the second article, Pennycook et al. (2018) investigated the impact of prior exposure on the identification of fake news. We will first discuss the reanalysis of Pennycook and Rand's (2019) data on the role of cognitive and motivational factors, before turning to the reanalysis of Pennycook et al.'s (2018) data on the effects of prior exposure. Although our reanalysis provides more nuanced insights into the effects of partisan bias, cognitive reflection, motivated reflection, and prior exposure, the purpose of our reanalysis goes beyond these insights, in that it aims to illustrate the broader value of SDT for research on the identification of fake news.

Lazy, Biased, or Both?

The main goal of Pennycook and Rand's (2019) studies was to investigate the role of cognitive and motivational factors in the identification of fake news. According to Pennycook and Rand (2019), cognitive and motivational accounts provide different explanations as to why people fall for fake news. Cognitive accounts suggest that people fall for fake news when they fail to engage in analytical thinking. In contrast, motivational accounts suggest that people fall

[4] From a cognitive perspective, prior exposure may influence response biases in two different ways. First, prior exposure may lower participants' decision threshold, in that they become more liberal in judging news headlines as real. Second, prior exposure may increase the perceived veracity of headlines with the decision threshold being unaffected.

for fake news because they are motivated to see the world in a particular way. Based on the two explanations, Pennycook and Rand (2019) derived competing predictions about the impact of analytical thinking—as measured by the Cognitive Reflection Test (Frederick, 2005)—on people's susceptibility to fake news. For cognitive accounts, the authors predicted that participants with higher CRT scores should be *less* susceptible to partisan fake news than participants with lower CRT scores, because participants with a greater propensity to engage in analytical thinking should be better at distinguishing real news from fake news. In contrast, for motivational accounts, the authors derived the prediction that participants with higher CRT scores should be *more* susceptible to partisan fake news than participants with lower CRT scores, because participants with a greater propensity to engage in analytical thinking should be better at strategically processing information in a manner such that the inferential outcomes are consistent with their cherished beliefs.

To test these competing predictions, Pennycook and Rand (2019) conducted two high-powered studies in which participants were asked to identify fake news in a set of news headlines. The set included both real news and fake news that were either pro-Republican or pro-Democrat. For each headline, participants were asked: "to the best of your knowledge, how accurate is the claim in the above headline" (Pennycook & Rand, 2019, p. 41). To investigate the role of cognitive and motivational factors, participants were asked to complete the CRT and a measure of political ideology. Across the two experiments, CRT scores were negatively correlated with the perceived accuracy of fake news and positively correlated with the ability to distinguish between real news and fake news. Moreover, the negative correlation between CRT scores and perceived accuracy of fake news was unrelated to the congruence of the headline with participants' political ideology. Based on these findings, the authors concluded that "susceptibility to fake news is driven more by lazy thinking than it is by partisan bias" (p. 39).

Our reanalysis of Pennycook and Rand's (2019) data using SDT suggests that the roles of cognitive reflection and partisan bias in the identification of fake news are more complex. Recall that, when analyzed from the perspective of SDT, effects of cognitive reflection and partisan bias are not mutually exclusive, because their respective effects pertain to different aspects (i.e., discrimination sensitivity vs. response bias). It is also worth noting that, based our conceptual analysis in terms of SDT, Pennycook and Rand's (2019) prediction for motivational accounts refers to effects of motivated reflection, not partisan bias per se. As explained above, a purely cognitive effect of reflection should be evident in discrimination sensitivity scores (d'), in that greater cognitive reflection should be associated with a stronger ability to distinguish real news and fake news. In contrast, partisan bias should be evident in response bias scores (c), in that people should show a lower threshold for judging news articles as real when they are congruent than when they are incongruent with their ideological beliefs. Finally, motivated reflection should lead to an interactive effect of cognitive reflection and ideology congruence on response bias scores (c), in that the tendency to accept ideology-congruent news as real and dismiss ideology-incongruent news as fake should be more pronounced for people with stronger reflective abilities. Thus, from the perspective of SDT, the outcomes predicted by the cognitive-reflection account, the partisan-bias account, and the motivated-reflection account are not mutually exclusive, as incorrectly implied by Pennycook and Rand's (2019) question of whether analytical thinking makes people more or less susceptible to fake news.

To gain deeper insights into the effects of cognitive reflection and ideology congruence on the identification of fake news, we reanalyzed Pennycook and Rand's (2019) data using SDT by calculating (1) *d′* scores reflecting participants' ability to accurately distinguish real news from fakes and (2) *c* scores reflecting participants' response bias in judging news as real versus fake regardless of news veracity. We calculated *d′* scores such that higher scores reflect greater accuracy in discriminating real news and fake news; *c* scores were calculated such that scores greater than zero reflect a response bias to judge headlines as fake and scores smaller than zero a response bias to judge headlines as real regardless of their veracity. To investigate the robustness of the obtained effects, we conducted SDT analyses for each of the two studies as well as an integrative data analysis (IDA) of the data from both studies (see Curran & Hussong, 2009). The details of our reanalysis are presented in Appendix A.

Consistent with Pennycook and Rand's (2019) conclusion, our reanalysis using *d′* scores indicates that participants' ability to discriminate between real news and fake news increased as a function of analytical thinking, as reflected in a significant positive association between CRT scores and *d′* scores (see Figure 3). This association was statistically significant in Study 1, Study 2, and the IDA (see Table 2). Moreover, participants were better in discriminating between real news and fake news when the headlines were congruent than when they were incongruent with their political ideology (see Figure 3). This difference was statistically significant in Study 1, Study 2, and the IDA (see Table 2). Our analysis also revealed evidence for an interaction between analytical thinking and ideology congruence, such that the positive association between CRT scores and accuracy in discriminating real news and fake news was stronger for politically congruent headlines than for politically incongruent headlines (see Figure 3). However, this interaction was statistically significant only in Study 1 and the IDA, but not in Study 2 (see Table

2).[5] Together, these findings suggest that people are better at distinguishing between real news and fake news when the content is congruent than when it is incongruent with their political ideology. Moreover, the ability to accurately distinguish between real news and fake news increases as a function of analytical thinking.

A major advantage of SDT is that it provides a tool to disentangle discrimination sensitivity and response biases. This distinction is particularly important for understanding the role of partisan bias and motivated reflection in the identification of fake news, because either of these factors should influence the identification of fake news via responses biases, not discrimination sensitivity. Thus, the fact that our reanalysis using d' scores supports the postulated role of cognitive reflection does not speak against the possibility that partisan bias and motivated reflection influence *c* scores in a manner predicted by extant accounts (i.e., disjunctive fallacy; see Gawronski & Bodenhausen, 2015).

Indeed, consistent with the proposed role of partisan bias, our analysis using *c* scores revealed that participants were more likely to judge politically incongruent headlines as a fake regardless of veracity compared to politically congruent headlines (see Figure 4). This difference was statistically significant in Study 1, Study 2, and the IDA (see Table 2). Interestingly, there was also evidence for a positive association between CRT scores and *c* scores, indicating that participants with a stronger propensity to engage in analytical thinking were more likely to dismiss all headlines as fake news regardless of veracity compared to participants with a weaker

[5] A potential interpretation of the obtained interaction between cognitive reflection and ideology congruence is that (1) analytical thinking supports accurate fake news discernment via enhanced engagement with political information and (2) effects of political engagement tend to be more pronounced for ideology-congruent than ideology-incongruent information due to selective exposure to ideology-congruent information in echo chambers. However, because the interaction between CRT and ideology-congruence was very small overall and not statistically significant in Study 2, we refrain from drawing strong conclusions from this effect.

propensity to engage in analytical thinking (see Figure 4). However, this association was statistically significant only in Study 2 and the IDA, but not in Study 1 (see Table 2). The interaction between CRT scores and ideology congruence was not significant in Study 1, Study 2, and the IDA (see Table 2). The latter finding speaks against the idea that analytical thinking increases partisan bias, as suggested by motivated-reflection accounts (see Pennycook & Rand, 2019). Nevertheless, the significant effect of ideology congruence suggests that partisan bias influences the identification of fake news via responses biases over and above the obtained effect of cognitive reflection on discrimination sensitivity. Interestingly, although higher cognitive reflection was associated with greater accuracy in distinguishing between real news and fake news, it did not reduce partisan bias.[6]

Together, our reanalysis of Pennycook and Rand's (2019) data using SDT offers a more nuanced picture. Different from their conclusion that "susceptibility to fake news is driven more by lazy thinking than it is by partisan bias" (p. 39), our analysis suggests that both factors can make people fall for fake news. On the one hand, "lazy thinking" can increase people's susceptibility to fake news by reducing their ability to distinguish real news from fake news. On the other hand, partisan bias can increase people's susceptibility to fake news by inducing a response bias to accept information that is congruent with their ideological beliefs and dismiss information that incongruent with their ideological beliefs regardless of veracity, and this bias seems to be unaffected by reflective thinking.

[6] An interesting secondary finding is that partisan bias in judgments of ideology-congruent and ideology-incongruent news headlines was more pronounced among self-identified Republicans than self-identified Democrats. This difference was statistically significant in Study 1, $t(796) = 4.70$, $p < .001$, $\eta_G^2 = .008$, Study 2, $t(2625) = 7.19$, $p < .001$, $\eta_G^2 = .005$, and the IDA, $t(3425) = 8.20$, $p < .001$, $\eta_G^2 = .005$.

Effects of Prior Exposure

The main goal of Pennycook et al.'s (2018) studies was to investigate the impact of prior exposure on fake news discernment. Research on the illusory truth effect suggests that prior exposure increases perceptions of veracity by increasing the fluency of processing the relevant information (Lewandowsky et al., 2012; Schwarz et al., 2007; Unkelbach et al., 2019). This effect seems highly relevant for the identification of fake news on social media, because echo chambers can increase the likelihood of multiple exposures to the same piece of misinformation (Schwarz & Jalbert, in press; Törnberg, 2018).

To investigate the emergence of illusory truth effects in the context of fake news, Pennycook et al. (2018) used a paradigm where participants first indicated for a set of real and fake news headlines whether or not they would share the story. Afterwards, participants were presented with the same fake and real news headlines from the prior task as well as novel fake and real news headlines that were not presented before. As in Pennycook and Rand (2019), participants were asked: "to the best of your knowledge, how accurate is the claim in the above headline?" (Pennycook et al., 2018, p. 1870). In one study, the manipulation of prior exposure and the measurement of perceived veracity occurred within the same session (Study 2). A follow-up study additionally measured perceived veracity one week later (Study 3). Thus, whereas in the former study the number of prior exposures could be 0 or 1, the number of prior exposures in the latter study could be 0, 1, or 2.[7] Based on prior research on the illusory truth effect, Pennycook et al. (2018) predicted that the likelihood for fake news headlines to be judged as real would increase as a result of prior exposure. Consistent with this hypothesis, participants

[7] The two studies also included a manipulation of explicit warnings about lack of veracity. Because this manipulation was not part of the main scope of Pennycook et al.'s (2018) original article, we did not include it in our reanalysis using SDT.

were more likely to judge fake news headlines as real when participants had been exposed to the headlines before than when they had not been exposed to the headlines before.

From the perspective of SDT, a potential interpretation of Pennycook et al.'s (2018) findings is that prior exposure influenced the identification of fake news via response biases, in that prior exposure to news headlines led to a tendency to judge these headlines as real regardless of their veracity. Yet, another possibility is that prior exposure influenced the identification of fake news via discrimination sensitivity, in that prior exposure reduced participants' ability to correctly distinguish real news from fake news.

To gain deeper insights into how prior exposure influences the identification of fake news, we reanalyzed Pennycook et al.'s (2018) data using SDT. Toward this end, we calculated d' scores in a manner such that higher scores reflect greater accuracy in discriminating real news and fake news; c scores were calculated in a manner such that scores greater than zero reflect a response bias to judge headlines as fake and scores smaller than zero a response bias to judge headlines as real regardless of their veracity. To investigate the robustness of the obtained effects, we again conducted SDT analyses for each of the two studies as well as an IDA of the data from both experiments (see Curran & Hussong, 2009). The details of our reanalysis are presented in Appendix B.

Consistent with the idea that prior exposure affected the identification of fake news via discrimination sensitivity, our reanalysis using d' scores indicates that participants' ability to discriminate between real news and fake news decreased as a function of prior exposure (see Figure 5). This conclusion is supported by a significant negative association between prior exposure and d' scores in Pennycook et al.'s (2018) Study 3 and the IDA (see Table 3). However, this association was not statistically significant in Pennycook et al.'s (2018) Study 2 (see Table 3).

Moreover, consistent with the idea that prior exposure affected the identification of fake news via responses biases, our reanalysis using *c* scores indicates that participants' tendency to dismiss news headlines as fake regardless of their veracity decreased as a function of prior exposure (see Figure 6). This conclusion is supported by a significant negative association between prior exposure and *c* scores in Pennycook et al.'s (2018) Study 2, Study 3, and the IDA (see Table 3).

Together, our reanalysis of Pennycook et al.'s (2018) data using SDT suggest that prior exposure can influence the identification of fake news in two functionally distinct ways. First, prior exposure may influence the identification of fake news by reducing people's ability to accurately discriminate between real news and fake news. Second, prior exposure may influence the identification of fake news by inducing a tendency to judge previously encountered news as real regardless of their actual veracity. These findings have important implications not only for applied research on the identification of fake news; they also provide valuable information for basic research on the mechanisms underlying the illusory truth effect (for a review, see Unkelbach et al., 2019).

Implications and Future Directions

The reported reanalyses demonstrate the value of SDT in providing more nuanced insights into how partisan bias, cognitive reflection, and prior exposure influence the identification of fake news. By distinguishing between discrimination sensitivity and response biases, our reanalysis revealed that ideological beliefs influenced judgments via a response bias to accept ideology-congruent news as real and dismiss ideology-incongruent news as fake regardless of news veracity. Nevertheless, cognitive reflection was found to be associated with veracity judgments in two distinct ways by (1) increasing overall accuracy in discriminating between real news and fake news (especially for ideology-congruent news) and (2) increasing response biases to judge news as fake regardless of veracity. There was no evidence for an effect of motivated reflection, in that partisan bias in the acceptance of ideology-congruent news and rejection of ideology-incongruent news did not increase as a function of cognitive reflection. Yet, cognitive reflection did not reduce partisan bias either, despite its positive association with the ability to accurately discriminate between real news and fake news. Finally, prior exposure was found to have a dual impact, in that it (1) reduced the ability to correctly distinguish between real news and fake news and (2) increased the likelihood that news is judged as real regardless of its veracity.

Although effects of partisan bias, cognitive reflection, motivated reflection, and prior exposure have received considerable attention in previous research on the identification of fake news, future research on other important factors may similarly benefit from SDT's capacity to disentangle discrimination sensitivity and response biases. One example is research on the effects of source-related information, especially information about the source's trustworthiness (see Kruglanski et al., 2005). At the most basic level, people may use the source of a news article as a cue to judge the credibility of the article's content, in that some known sources might be perceived as more trustworthy than others (e.g., *Wall Street Journal* vs. *National Enquirer*). In addition, people may be more skeptical about the trustworthiness of unknown sources compared to known reputable sources (see Schwarz & Jalbert, 2020). Although using source-related information as a cue for credibility may be a valuable heuristic when navigating through the massive amount of real and fake news on social media, it is worth noting that higher levels of context-specific accuracy associated with this heuristic in a particular environment should not be confused with overall discrimination sensitivity in terms of SDT. After all, it seems likely that people accept information from sources they trust and dismiss information from sources they do not trust regardless of the information's actual veracity (Pilditch, Madsen, & Custers, 2020). From the perspective of SDT, source credibility may influence the identification of fake news via responses biases, but it may not necessarily increase people's ability to accurately distinguish between real news and fake news based on information content (e.g., correct discrimination of real news and fake news based on independent evidence; see Schwarz & Jalbert, 2020).

Potential effects of source-related information can be even more complex, considering that people may systematically differ in their perceptions of trustworthy and untrustworthy sources. For example, whereas Democrats may perceive CNN as a more trustworthy source of political information than FOX News, Republicans may have the opposite perception. To the extent that source credibility influences veracity judgments via response biases, this possibility suggests a second layer of partisan bias that goes beyond the asymmetric acceptance of ideology-congruent versus ideology-incongruent fake news (Van Bavel & Pereira, 2018). Using SDT to disentangle discrimination sensitivity and response biases may help to provide deeper insights into how source-related information influences the identification of fake news.

A related question with important implications for potential interventions is how people could be trained to improve their skills in detecting fake news. A recent study with close to 8,000 participants from 12 states in the United States found that a substantial proportion of students from middle school to college showed rather poor performance in distinguishing real news from fake news on the internet (Wineburg, McGrew, Breakstone, & Ortega, 2016). Such findings echo calls for interventions to increase students' digital literacy early in high school (e.g., McGrew, Smith, Breakstone, Ortega, & Windeburg, 2019). Yet, when evaluating the effectiveness of any such interventions, it seems important to distinguish between discrimination sensitivity and responses biases. From the perspective of SDT, interventions that improve people's ability to detect fake news may do so either (1) by increasing people's ability to correctly discriminate between real news and fake news or (2) by increasing responses bias to dismiss news as fake regardless of news veracity (or both). The possibility of such multifaceted effects can be illustrated with the findings of our reanalyses, suggesting that cognitive reflection is associated with both (1) greater accuracy in distinguishing between real news and fake news and (2) a greater response bias to dismiss news as fake regardless of news veracity. Although the latter effect resonates with the idea that a healthy dose of skepticism might buffer unwanted effects of misinformation (Lewandowsky et al., 2012), interventions that increase people's accuracy in discriminating between real news and fake news would seem more desirable compared to interventions that merely increase people's general distrust in the news media. The latter effect could be particularly problematic if the resulting skepticism is greater for ideology-incongruent than ideology-congruent information, as suggested by research on motivated skepticism (Ditto & Lopez, 1992).

Another interesting question for future research is how the processes underlying partisan bias in the identification of fake news might immunize people to the dismissed contents of ideology-incongruent news. A considerable body of research suggests that misinformation continues to impact judgments and decisions even after being refuted (Lewandowsky et al., 2012; Rapp & Braasch, 2014; Schwarz et al., 2007; for a meta-analysis, see Chan et al., 2017). However, this well-established finding seems to conflict with the anecdotal idea that people tend to be rather immune to the contents of real news they dismiss as fake. To the extent that the latter idea can be supported by empirical data, it would conflict with the vast amount of evidence for the relative ineffectiveness of invalidation and debunking. Yet, the resulting paradox would raise the interesting possibility that there is something distinct about the mechanisms underlying partisan bias in the identification of fake news that makes these mechanisms more effective in preventing effects of “invalidated” information. Research identifying these distinct features could provide valuable insights for improving the effectiveness of fact checking and the debunking of misinformation. SDT would be a valuable tool in this endeavor for its capacity to provide more nuanced insights into the determinants of discrimination sensitivity and response biases.

Another valuable aspect of adopting an SDT framework in research on the identification of fake news is that it provides conceptual links to other areas that may inform broader theorizing on judgment and decision-making. In the introduction, we already mentioned research on recognition memory (Snodgrass & Corwin, 1988) and racial bias in weapon identification (Payne & Correll, 2020). Other examples are studies that have used SDT to quantify discrimination sensitivity and responses biases in the illusory truth effect (e.g., Unkelbach, 2007) and eyewitness identification (e.g., Wixted, Mickes, Dunn, Clark, & Wells, 2016). In the latter line of work, SDT has provided valuable insights into differences between sequential and simultaneous lineups. Based on findings suggesting that innocent “fillers” are less frequently identified as suspects in sequential lineups compared to simultaneous lineups, some researchers concluded that sequential lineups are diagnostically superior (Steblay, Dysart, & Wells, 2011). However, SDT analyses suggest that the decrease in incorrect identifications is due to the impact of lineup type on response bias, not discrimination sensitivity (Wixted et al., 2016). That is, people are not more accurate in sequential lineups; they are simply more conservative. If anything, the available evidence suggests that sequential lineups reduce discrimination sensitivity (Mickes & Wixted, in press). An SDT framework not only avoids such misinterpretations of classification results (see also Dube, Rotello, & Heit, 2010); it also helps to organize findings in a given area. For example, in a recent review of research on truth evaluation, Brashier and Marsh (2020) have used SDT to organize the available evidence, describing the impact of knowledge on discrimination sensitivity and the impact of credulity on response bias in judgments of truth. As research on fake news detection grows (Greifeneder, Jaffé, Newman, & Schwarz, 2020; Rapp & Braasch, 2014), an SDT framework may prove similarly helpful in organizing the available evidence, providing valuable links for broader theorizing on judgment and decision-making.

Some Caveats

The main goal of the current work was to illustrate the value of SDT in providing more nuanced insights into the processes underlying the identification of fake news. Yet, to avoid potentially premature conclusions, it seems appropriate to mention a few caveats. First, it is worth noting that the sample sizes of the reanalyzed data sets were quite large. Although large sample sizes have the advantage of reducing the likelihood of both false positives (Button et al., 2013) and false negatives (Maxwell, Lau, & Howard, 2017), they also increase statistical power for the detection of very small effects that may be negligible from a practical point of view (Wilson, Harris, & Wixted, 2020). In terms of current conventions regarding the interpretation of effect sizes (Cohen, 1988), the only effect that was close to medium-size level was the obtained pattern of partisan bias in the acceptance of ideology-congruent news and the rejection of ideology-incongruent news (see Appendix A, Table 2). Some of the obtained effects qualify as small in terms of current conventions, including the association between cognitive reflection and discrimination sensitivity (see Appendix A, Table 2), the effect of ideology congruence on discrimination sensitivity (see Appendix A, Table 2), and the effect of prior exposure on response bias (see Appendix B, Table 3). Yet, other effects fall below the conventional benchmark for small effects, including the association between cognitive reflection and response bias (see Appendix A, Table 2), the interactive effect of cognitive reflection and ideology congruence on discrimination sensitivity (see Appendix A, Table 2), and the effect of prior exposure on discrimination sensitivity (see Appendix B, Table 3). Thus, although our reanalysis illustrates the relation between seemingly conflicting hypotheses and the value of SDT in providing more nuanced insights into the processes underlying the propagation of fake news, the practical importance of these findings may better be evaluated in terms of the obtained effect sizes. Moreover, because the number of real and fake news headlines was very small in both Pennycook and Rand's (2019) and Pennycook et al.'s (2018) studies, and because small stimulus sets can distort statistical results (Judd, Westfall, & Kenny, 2012), substantive conclusions from the reported findings would benefit from follow-up studies with larger stimulus sets. Although these considerations give reasons to be cautious in the conclusions that may be drawn from the obtained results, they do not qualify our central point: the value of SDT in disentangling different aspects in the identification of fake news.

Another caveat concerns the dominant emphasis on accuracy judgments in studies on the identification of fake news (e.g., Pennycook & Rand, 2019; Pennycook et al., 2018), which may not reflect the mindset with which people process news outside the lab. Indeed, some researchers have argued that identity-related motivations may override accuracy motivation in most real-world settings (e.g., Van Bavel & Perreira, 2018), raising important questions about whether the effects obtained for accuracy judgments generalize to other important decisions, such as decisions to share news on social media. In line with this concern, people seem to be willing to share repeatedly encountered misinformation even when they are aware that the information is factually incorrect (Effron & Raj, 2020). Although our reanalyses focused primarily on judgments of veracity, SDT can also be applied to analyze sharing decisions, with *d′* reflecting the tendency to share real news and not share fake news and *c* reflecting the tendency to share (vs. not share) news regardless of veracity. Based on the concern that veracity judgments may not reflect effects of identity-related motivations that guide sharing decisions in real-world contexts, future research using SDT to study effects of partisan bias, cognitive reflection, and prior exposure on sharing decisions would be helpful to evaluate the generality of the obtained results.

From a technical view, it also seems appropriate to acknowledge alternatives to SDT that would accomplish the goal of disentangling sensitivity and bias in the identification of fake news (e.g., high-threshold model, process dissociation procedure). Each of these alternatives is based on different assumptions about the mechanisms underlying detection (e.g., high-threshold model would assume a headline is either detected as a fake news or not, with no nuance in between; see Blackwell, 1953), the characteristics of perceived accuracy distributions for the two kinds of stimuli (e.g., Gaussian distributions of equal vs. unequal variance in SDT; see Green & Swets, 1966; Wixted, 2020), and the relation between the two aspects (e.g., bias being conditional upon the absence of sensitivity in process dissociation; see Jacoby, 1991). Although we deem SDT superior to extant alternatives for research on the identification of fake news, we cannot rule out that alternative models may be more appropriate for this endeavor. Yet, regardless of the preferred approach, we deem it essential to make the background assumptions of the utilized model explicit. This concern applies even to seemingly “atheoretical” approaches, such as using the raw percentage of news identified as fake, which is based on conceptual background assumptions of high-threshold models (e.g., headlines are identified as either real or fake, with no nuance in between; see Wixted, 2020). Future research could resolve these ambiguities by including measures of confidence, which allows direct tests of different background assumptions by means receiver operating characteristic (ROC) analyses.

Conclusion

The main goal of the current article was to illustrate the value of SDT in providing more nuanced insights into the determinants of fake new beliefs. The most significant feature of SDT is its capacity to disentangle two conceptually distinct aspects in the identification of fake news: (1) ability to correctly distinguish between real news and fake news and (2) response biases to judge news as real versus fake regardless of news veracity. Although SDT was developed more than 50 years ago and has been applied to a wide range of topics within psychology, extant research on the identification of fake news has not yet utilized the beneficial features of SDT. We hope that the insights offered by our reanalyses of existing data will inspire researchers in this area to adopt SDT in their own work, providing a better understanding of why people fall for fake news.

Acknowledgment

We thank Gordon Pennycook for making the data sets of Pennycook and Rand’s (2019) and Pennycook et al.’s (2018) studies publicly accessible, and for his invaluable help with the code book for the data sets.

References

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General, 149,* 1608-1613.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*. Retrieved from https://doi.org/10.18637/jss.v067.i01

Blackwell, H. R. (1953). Psychophysical thresholds: Experimental studies of methods of measurement. *University of Michigan Research Institute Bulletin*, *36*. Ann Arbor: University of Michigan.

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

Bryan, J. (2019). *Googlesheets4: Access Google Sheets using the Sheets API V4*. Retrieved from https://CRAN.R-project.org/package=googlesheets4

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 1-12.

Chan, M.-P., Jones, C. R., Jamieson, K. H., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science, 28,* 1531-1546.

Conroy, N. J., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology, 52,* 1-4.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14,* 81-100.

Ditto, P. H., & Lopez, D. F. (1992). Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology, 63*, 568-584.

Dube C., Rotello C. M., & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review, 117,* 831–863.

Effron, D. A., & Raj, M. (2020). Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science*, *31*, 75-87.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19,* 25-42.

Gawronski, B. (2012). Back to the future of dissonance theory: Cognitive consistency as a core motive. *Social Cognition, 30,* 652-668.

Gawronski, B., & Bodenhausen, G. V. (2015). Theory evaluation. In B. Gawronski, & G. V. Bodenhausen (Eds.), *Theory and explanation in social psychology* (pp. 3-23). New York: Guilford Press.

Greifeneder, R., Jaffé, M., Newman, E. J., & Schwarz, N. (Eds.). (2020). *The psychology of fake news: Accepting, sharing, and correcting misinformation.* London: Routledge

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers*, *27*, 46-51.

Hester, J. (2019). *Glue: Interpreted String Literals*. Retrieved from https://github.com/tidyverse/glue

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513-541.

Jaeger, B. C., Edwards, L. J., Das, K., & Sen, P. K. (2017). An R^2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, *44*, 1086-1105.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review, 111,* 640-661.

Judd, C. M., Kenny, D. A., & McClelland, G. H. (2001). Estimating and testing mediation and moderation in within-subject designs. *Psychological Methods*, *6*, 115-134.

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). *Data analysis: A model comparison approach to regression, ANOVA, and beyond* (3rd Edition). New York: Taylor & Francis.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor : Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68,* 601-628.

Kahan, D. M. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition. *Cultural Cognition Project Working Paper Series No. 164*. Retrieved from https://ssrn.com/abstract=2973067

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioral Public Policy, 1,* 54-86.

Kruglanski, A., Raviv, A., Bar-Tal, D., Raviv, A., Sharvit, K., Ellis, S., Bar, R., Pierro, A., & Mannetti, L., (2005). Says who? Epistemic authority effects in social judgment. *Advances in Experimental Social Psychology, 37,* 345-392.

Kruglanski, A., & Webster, D. M. (1996). Motivated closing of the mind: “Seizing and freezing.” *Psychological Review, 103,* 263-283.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108,* 480-498.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., …, & Zittrain, J. L. (2018). The science of fake news: Addressing fake news requires a multidisciplinary effort. *Science, 359,* 1094-1096.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13,* 106-131.

Macmillan, N. A., & Creeman, C. D. (2004). *Detection theory: A user’s guide.* New York: Taylor and Francis.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist, 70,* 487-498.

McGrew, S., Smith, M., Breakstone, J., Ortega, T., & Windeburg, S. (2019). Improving university students’ web savvy: An intervention study. *British Journal of Educational Psychology, 89,* 485-500.

Mickes, L. & Wixted, J. T. (in press). Eyewitness memory. In M. J. Kahana & A. D. Wagner (Eds.), *The Oxford Handbook of human memory*. New York: Oxford University Press

Mitchell, A., Gottfried, J., Stocking, G., Walker, M., & Fedeli, S. (2019). *Many Americans say made-up news is a critical problem that needs to be fixed.* Pew Research Center Report. Retrieved from https://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/

Payne, B. K., & Correll, J. (2020). Race, weapons, and the perception of threat. *Advances in Experimental Social Psychology, 62,* 1-50.

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases the perceived accuracy of fake news. *Journal of Experimental Psychology: General, 147,* 1865-1880.

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

Pilditch, T. D., Madsen, J. K., & Custers, R. (2020). False prophets and Cassandra's curse: the role of credibility in belief updating. *Acta Psychologica, 202:102956.*

Rapp, D. N., & Braasch, J. L. G. (2014). *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. Cambridge, MA: MIT Press.

Reber, R., & Unkelbach, C. (2010). The epistemic status of fluency as source for judgments of truth. *Review of Philosophy and Psychology, 1,* 563-581.

Rudis, B. (2019). *Hrbrthemes: Additional themes, theme components and utilities for "ggplot2".* Retrieved from https://CRAN.R-project.org/package=hrbrthemes

Schwarz, N., & Jalbert, M. (2020). When (fake) news feels true: Intuitions of truth and the acceptance and correction of misinformation. R. Greifeneder, M. Jaffé, E. J. Newman, & N. Schwarz (Eds.), *The psychology of fake news: Accepting, sharing, and correcting misinformation* (pp. 73-90). London: Routledge.

Schwarz, N., Sanna, L., Skurnik, I., & Yoon, C. (2007). Metacognitive experiences and the intricacies of setting people straight: Implications for debiasing and public information campaigns. *Advances in Experimental Social Psychology, 39,* 127-161.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). *Afex: Analysis of* *factorial experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117,* 34-50.

Stanislav, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers, 31,* 137-149.

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy and Law, 17,* 99–139.

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PloS One, 13(9):e0203958.*

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 219–230.

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science, 28,* 247-253.

Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences, 22,* 213-224.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4(43):1686.*

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences, 117,* 5559-5567

Wineburg, S., McGrew, S., Breakstone, J., & Ortega, T. (2016). *Evaluating information: The cornerstone of civic online reasoning*. Stanford Digital Repository. Retrieved from http://purl.stanford.edu/fv751yt5934

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 46,* 201-233.

Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, *121*, 262-276.

Wixted, J. T., Mickes, L., Dunn, J. C., Clark, S. E., & Wells, W. (2016). Estimating the reliability of eyewitness identifications from police lineups. *Proceedings of the National Academy of Sciences, 113,* 304-309.

Appendix A: SDT Analysis of Data from Pennycook and Rand (2019)

Our SDT analysis of data from Pennycook and Rand (2019) is based on the publicly available materials provided by the authors at https://osf.io/tuw89/. All materials for the current analysis (i.e., data wrangling, data analysis, reporting R scripts) are publicly available at https://osf.io/uc9me/?view_only=2cb2a4d6f0df4ef1abbd01e1d5b58674.[8] We used (among others), the following R packages: afex (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2020), glue (Hester, 2019), hrbrthemes (Rudis, 2019), the tidyverse (Wickham et al., 2019).

Data preparation. We computed two *d′* and two *c* indices for each participant, one for politically congruent headlines and one for politically incongruent headlines.[9] To compute these indices, we used the $d' = z(H) - z(FA)$ and $c = ½ [(z(H) + z(FA)]$ formulas, with z(X) depicting the quantile function for z distribution such that a proportion of 0.5 is converted to a z-score of 0 (Stanislaw & Todorov, 1999). *H* refers to the proportion of real news articles that were judged as real (i.e., hit rate) and *FA* refers to the proportion of fake news articles that were judged as real (i.e., false alarm rate). Because of the low number of trials per condition, we used Hautus's (1995) corrections for *d′* and *c*. We calculated *d′* scores such that higher scores reflect greater accuracy in discriminating real news and fake news; *c* scores were calculated such that scores greater than zero reflect a response bias to judge headlines as fake and scores smaller than zero a response bias to judge headlines as real regardless of their veracity.

[8] R files starting with a 0 are the ones used for the current reanalysis (i.e., 000_data-wrangling.R, 001_SDT.R). Original data, as downloaded from https://osf.io/tuw89/, can be found in the data-original folder. Data sets used for the analysis can be found in the data-raw folder. Post-wrangling data sets can be found in the data-tidy folder.

[9] Technically, the four response options on the measure of perceived veracity would have provided two pairs of indices with different levels of confidence. Such data would provide a basis for ROC analyses, which can be informative regarding the model underlying signal detection (e.g., equal-variance vs. unequal-variance; see Wixted, 2020). However, because of the small number of observations for each participant, it was not possible to compute indices at different levels of confidence. We therefore dichotomized judgments of perceived accuracy.

Analyses. We adopted a model comparison approach to predict d' and c, respectively (see Judd, McClelland, & Ryan, 2017). For each study, we predicted the two SDT scores by the ideological congruency of the headline as a within-subjects factor and CRT scores as a between-subject factor.[10] Political congruency was contrast-coded, such that ideology-congruent headline corresponded to -1 and ideology-incongruent headline corresponded to 1. To investigate the robustness of the obtained effects, we ran this analysis separately for Study 1 and Study 2, followed by an IDA of the data from both studies (see Curran & Hussong, 2009). A summary of the results can be found in Table 2.

Study 1. The analysis for discrimination sensitivity revealed that average d' scores were significantly greater than zero, indicating that participants' ability to correctly distinguish between real news and fake news was above chance overall, $t(798) = 49.31$, $p < .001$, $\eta_G^2 = .667$. A significant positive association with CRT scores indicated that participants with high CRT scores were better at discriminating real news and fake news than participants with low CRT scores, $t(798) = 6.88$, $p < .001$, $\eta_G^2 = .037$. There was also a significant main effect of ideological congruency, indicating that participants were better at discriminating real news and fake news for ideology-congruent headlines than ideology-incongruent headlines, $t(798) = -10.38$, $p < .001$, $\eta_G^2 = .044$. These effects were qualified by a significant interaction between CRT and political congruency, indicating that the association between CRT and discrimination sensitivity was

[10] With a model comparison approach, as in a mixed-effect ANOVA, two linear regressions are used to estimate the effect of CRT and partisanship. For example, the following models are used for the d' analysis:

$$\frac{d'_{politically\ congruent\ i} + d'_{politically\ incongruent\ i}}{2} = b_{10} + b_{11}CRT_i + e_i$$

$$d'_{politically\ congruent\ i} - d'_{politically\ incongruent\ i} = b_{20} + b_{21}CRT_i + e_i$$

with b_{10} estimating the intercept of d', b_{11} the effect of CRT on d', b_{20} the effect of the headlines' partisanship on d', and $b_{21}$the interaction effect of CRT and headlines' partisanship on d'.

stronger for politically congruent than politically incongruent headlines, *t*(798) = - 3.42, $p < .001$, $\eta_G^2 = .005$.

The analysis for response bias revealed that average *c* scores were significantly greater than zero, indicating that participants had an overall tendency to judge the news headlines as fake regardless of their veracity, *t*(798) = 16.87, $p < .001$, $\eta_G^2 = .200$. There was no significant association with CRT scores, *t*(798) = 0.72, $p = .472$, $\eta_G^2 < .001$, but the main effect of political congruency was statistically significant, *t*(798) = 16.09, $p < .001$, $\eta_G^2 = .088$. The latter effect indicates that participants were more likely to judge a headline as a fake when it was incongruent than when it was congruent with their political ideology. The interaction between CRT and political congruency was not significant, *t*(798) = 0.36, $p = .722$, $\eta_G^2 < .001$.

Study 2. The analysis for discrimination sensitivity revealed that average d' scores were significantly greater than zero, indicating that participants' ability to correctly distinguish between real news and fake news was above chance overall, *t*(2627) = 83.64, $p < .001$, $\eta_G^2 = .632$. Replicating the findings of Study 1, there was a significant positive association with CRT scores, indicating that participants with high CRT scores were better at discriminating real news and fake news than participants with low CRT scores,*t*(2627) = 9.96, $p < .001$, $\eta_G^2 = .022$. Also replicating the findings of Study 1, a significant main effect of ideological congruency indicated that participants were better at discriminating real news and fake news for ideology-congruent headlines than ideology-incongruent headlines, *t*(2627) = -5.40, $p < .001$, $\eta_G^2 = .004$. The interaction between CRT and political congruency was not significant, *t*(2627) = -1.53, $p = .125$, $\eta_G^2 < .001$.

The analysis for response bias revealed that average *c* scores were significantly greater than zero, indicating that participants had a tendency to judge the news headlines as fake regardless of their veracity, $t(2627) = 35.94$, $p < .001$, $\eta^2_G = .271$. Replicating the findings of Study 1, a significant main effect of political congruency indicated that participants were more likely to judge a headline as a fake when it was incongruent than when it was congruent with their political ideology, $t(2627) = 17.50$, $p < .001$, $\eta^2_G = .028$. Yet, different from Study 1, there was also a significant association with CRT scores, $t(2627) = 4.59$, $p < .001$, $\eta^2_G = .006$, indicating that participants with high CRT scores had a stronger tendency to judge the news headlines as fake regardless of their veracity compared to participants with low CRT scores. The interaction between CRT and ideological congruency was not significant, $t(2627) = 0.10$, $p = .924$, $\eta^2_G < .001$.

IDA. The IDA of the combined data from the two studies revealed that average d' scores were significantly greater than zero, indicating that participants' ability to correctly distinguish between real news and fake news was above chance overall, $t(3427) = 97.70$, $p < .001$, $\eta^2_G = .638$. A significant association with CRT scores indicated that participants with high CRT scores were better at discriminating real news and fake news than participants with low CRT scores, $t(3427) = 11.86$, $p < .001$, $\eta^2_G = .026$. Moreover, participants were significantly better at discriminating real news and fake news when the headline was congruent than when it was incongruent with their political ideology, $t(3427) = - 9.78$, $p < .001$, $\eta^2_G = .010$. These effects were qualified by a significant interaction between CRT and ideological congruency, indicating that the association between CRT and discrimination sensitivity was stronger for ideologically congruent than ideologically incongruent headlines, $t(3427) = -3.24$, $p = .001$, $\eta^2_G = .001$.

Regarding *c* scores, the IDA revealed that participants had an overall tendency to judge the news headlines as fake regardless of their veracity, $t(3427) = 39.57$, $p < .001$, $\eta^2_G = .252$. As in Study 2, there was a significant association with CRT, indicating that participants with high

CRT scores had a stronger tendency to judge the headlines as fake regardless of their veracity compared to participants with low CRT scores, (3427) = 4.26, *p* < .001, η_G^2 = .004. A significant main effect of ideological congruency further indicated that participants were more likely to judge a headline as a fake when it was incongruent than when it was congruent with their political ideology, *t*(3427) = 23.22, *p* < .001, η_G^2 = .040. The interaction between CRT and ideological congruency was not significant, *t*(3427) = 0.51, *p* = .610, η_G^2 < .001.

Appendix B: SDT Analysis of Data from Pennycook, Cannon, and Rand (2018)

Our SDT analysis of data from Pennycook et al. (2018) is based on the publicly available materials provided by the authors at https://osf.io/txf46/. All materials for the current analysis (i.e., data wrangling, data analysis, reporting R scripts) are publicly available at https://osf.io/uc9me/?view_only=2cb2a4d6f0df4ef1abbd01e1d5b58674.[11] We used (among others), the following R packages: afex (Singmann et al., 2020), glue (Hester, 2019), googlesheets4 (Bryan, 2019), hrbrthemes (Rudis, 2019), lme4 (Bates et al., 2015), and the tidyverse (Wickham et al., 2019). The data preparation followed the procedures described in Appendix A.

Analyses. We again adopted a model comparison approach to predict d' and c, respectively. For each study, we predicted d' and c by the number of exposures.[12] To investigate the robustness of the obtained effects, we ran this analysis separately for Study 2 and Study 3, followed by an IDA of the data from both studies (see Curran & Hussong, 2009). Because number of exposures included three levels in Study 3, we analyzed effects of prior exposures following recommendations by Judd, Kenny, and McClelland (2001). To account for the discrepant number of exposures in Studies 2 and 3, we adopted a mixed-model approach in the IDA (see Judd et al., 2017), with the two SDT indices as DVs, number of exposures as continuous predictor, and participants as random factor. $R^2_{\beta*}$ is reported as effect size (see Jaeger, Edwards, Das, & Sen, 2017). A summary of the results can be found in Table 3.

[11] R files starting with a 1 are the ones used for the reported reanalysis (i.e., 100_data-wrangling.R, 101_SDT.R). Original data, as downloaded from https://osf.io/txf46/, can be found in the data-original folder. Data sets used for the analysis can be found in the data-raw folder. Post-wrangling data sets can be found in the data-tidy folder.

[12] Because of the small number of observations, it was not possible to compute indices at different levels of confidence (see Footnote 8). We therefore dichotomized judgments of perceived accuracy.

Study 2. The analysis for discrimination sensitivity revealed that average d' scores were significantly greater than zero, indicating that participants' ability to correctly distinguish between real news and fake news was above chance overall, $t(946) = 47.59, p < .001, \eta_G^2 = .638$. Number of prior exposures had no significant effect on participants' ability to discriminate between real news and fake news, $t(946) = - 0.24, p = .809, \eta_G^2 < .001$.

The analysis for response bias revealed that average *c* scores were significantly greater than zero, indicating that participants had a tendency to judge the news headlines as fake regardless of their veracity, $t(946) = 18.61, p < .001, \eta_G^2 = .216$. The effect of prior exposure was statistically significant, indicating that the tendency to judge the news headlines as fake regardless of their veracity was weaker for headlines that had been presented before compared to headlines that had not been presented before, $t(946) = - 8.47, p < .001, \eta_G^2 = .018$.

Study 3. The analysis for discrimination sensitivity revealed that average d' scores were significantly greater than zero, indicating that participants' ability to correctly distinguish between real news and fake news was above chance overall, $t(564) = 40.72, p < .001, \eta_G^2 = .647$. A significant effect of number of prior exposures indicated that discrimination sensitivity decreased as a function of prior exposures, $t(564) = - 1.96, p = 050, \eta_G^2 = .003$.

The analysis for *c* scores revealed that participants had a tendency to judge the news headlines as fake regardless of their veracity, $t(564) = 9.08, p < .001, \eta_G^2 = .085$. A significant effect of number of prior exposures indicated that the tendency to judge the news headlines as fake regardless of their veracity decreased as a function of prior exposures, $t(564) = - 5.91, p < .001, \eta_G^2 = .022$.

IDA. The analysis for discrimination sensitivity revealed that average d' scores were significantly greater than zero, indicating that participants' ability to correctly distinguish between real news and fake news was above chance level, t(2667.83) = 56.40, p < .001. As in Study 3, prior exposure had a significant effect on participants' ability to accurately discriminate between real news and fake news, in that discrimination sensitivity decreased as a function of the number of prior exposures, t(2271.25) = - 2.10, p < .001, $R^2_{\beta*}$ = .001.

For the bias parameter c, the analysis revealed that participants had a general tendency to judge the news headlines as fake regardless of their veracity, t(2538.05) = 23.40, p < .001. Moreover, as in Studies 2 and 3, the effect of prior exposure was statistically significant, indicating that the tendency to judge the news headlines as fake regardless of their veracity decreased as a function of the number of prior exposures, t(2239.33) = - 10.95, p < .001, $R^2_{\beta*}$ = .023.

Table 1. Signal Detection Theory uses hit and false-alarm rates to compute *d'*, a discrimination sensitivity index reflecting people's ability in distinguishing target stimuli (e.g., real news) from distracter stimuli (e.g., fake news), and *c*, a response bias index reflecting the threshold for judging stimuli as belonging to the category of target stimuli.

	Response "Target" (e.g., response "real")	Response "Distracter" (e.g., response "fake")
Target Stimuli (e.g., real news)	HIT	MISS
Distracter Stimuli (e.g., fake news)	FALSE ALARM	CORRECT REJECTION

Table 2. Summary statistics of SDT reanalysis of data by Pennycook and Rand (2019), predicting discrimination sensitivity (*d′*) and response bias (*c*) in the identification of fake news by CRT scores and congruency of the headline with participants' political ideology.

Study	**Index**	**Term**	**df**	***t***	***p***	η^2_G
1	*d′*	Intercept	798	49.31	< .001	.667
		CRT	798	6.88	< .001	.037
		Congruency	798	- 10.38	< .001	.044
		CRT × Congruency	798	- 3.42	< .001	.005
	c	Intercept	798	16.87	< .001	.200
		CRT	798	0.72	.472	< .001
		Congruency	798	16.09	< .001	.088
		CRT × Congruency	798	0.36	.722	< .001
2	*d′*	Intercept	2627	83.64	< .001	.632
		CRT	2627	9.56	< .001	.022
		Congruency	2627	- 5.40	< .001	.004
		CRT × Congruency	2627	- 1.53	.125	< .001
	c	Intercept	2627	35.94	< .001	.271
		CRT	2627	4.59	< .001	.006
		Congruency	2627	17.50	< .001	.028
		CRT × Congruency	2627	0.10	.924	< .001
IDA	*d′*	Intercept	3427	97.70	< .001	.638
		CRT	3427	11.86	< .001	.026
		Congruency	3427	- 9.78	< .001	.010
		CRT × Congruency	3427	- 3.24	.001	.001
	c	Intercept	3427	39.57	< .001	.252
		CRT	3427	4.26	< .001	.004
		Congruency	3427	23.22	< .001	.040
		CRT × Congruency	3427	0.51	.610	< .001

Note. CRT = Cognitive Reflection Test. SDT = Signal Detection Theory. IDA = Integrative Data Analysis.

Table 3. Summary statistics of SDT reanalysis of data by Pennycook, Cannon, and Rand (2018), predicting discrimination sensitivity (*d′*) and response bias (*c*) in the identification of fake news by prior exposure.

Study	**Index**	**Term**	**df**	***t***	***p***	η^2_G	$R^2_{\beta*}$
2	*d′*	Intercept	946	47.59	< .001	.638	-
		Exposure	946	- 0.24	.809	< .001	-
	c	Intercept	946	18.61	< .001	.216	-
		Exposure	946	- 8.47	< .001	.018	-
3	*d′*	Intercept	564	40.72	< .001	.647	-
		Exposure	564	- 1.96	.050	.003	-
	c	Intercept	564	9.08	< .001	.085	-
		Exposure	564	- 5.91	< .001	.022	-
IDA	*d′*	Intercept	2667.83	56.40	< .001	-	-
		Exposure	2271.25	- 2.10	< .001	-	.001
	c	Intercept	2538.05	23.40	< .001	-	-
		Exposure	2239.33	- 10.95	< .001	-	.023

Note. SDT = Signal Detection Theory. IDA = Integrative Data Analysis.

Figure 1. Graphical depiction of SDT's index for discrimination sensitivity (*d′*), reflecting the distance between the distributions of judgments about real and fake news along the judgmental dimension of veracity. Distributions that are as closer together along the judgment-relevant dimension have a lower *d′*, indicating that participants' ability in correctly discriminating between real news and fake news is relatively low (left panel). Distributions that are further apart along the perceived veracity dimension have a higher *d′*, indicating that participants' ability in correctly discriminating between real news and fake news is relatively high (right panel).

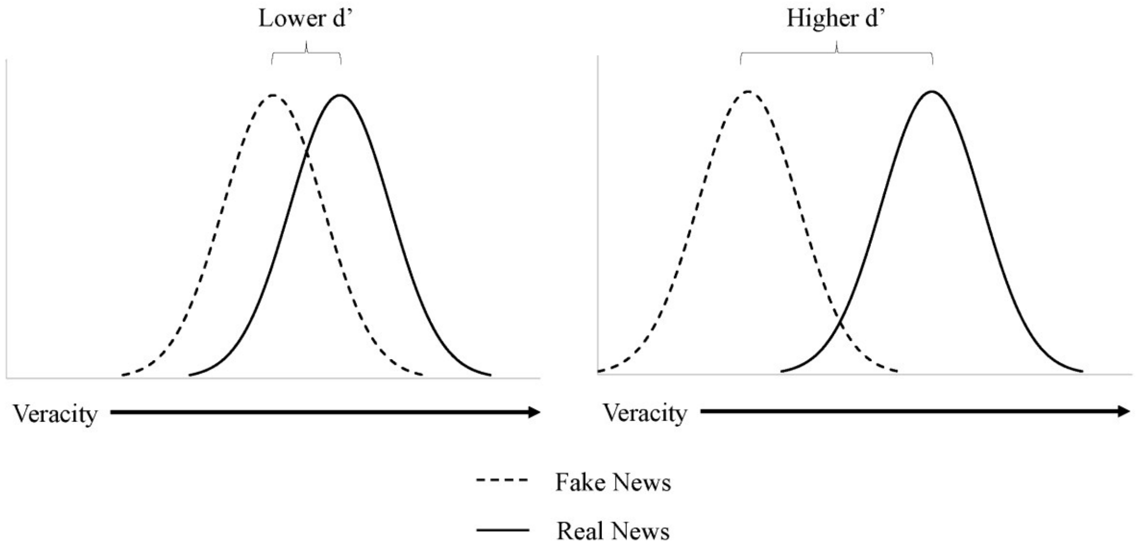


Figure 2. Graphical depiction of SDT's index for response bias (*c*), reflecting the threshold along the judgmental dimension of perceived veracity at which a participant decides to switch their decision. When judging whether news articles are real (vs. fake), *c* indicates the degree of veracity the participant must perceive before judging a news article as real. Any stimulus with greater perceived veracity than that value will be judged as real, whereas any stimulus with lower perceived veracity than that value will be judged as fake. A higher (or more conservative) threshold would indicate that a participant is generally less likely to judge a news story as real, whereas a lower (or more liberal) threshold would indicate that a participant is generally more likely to judge a news story as real.

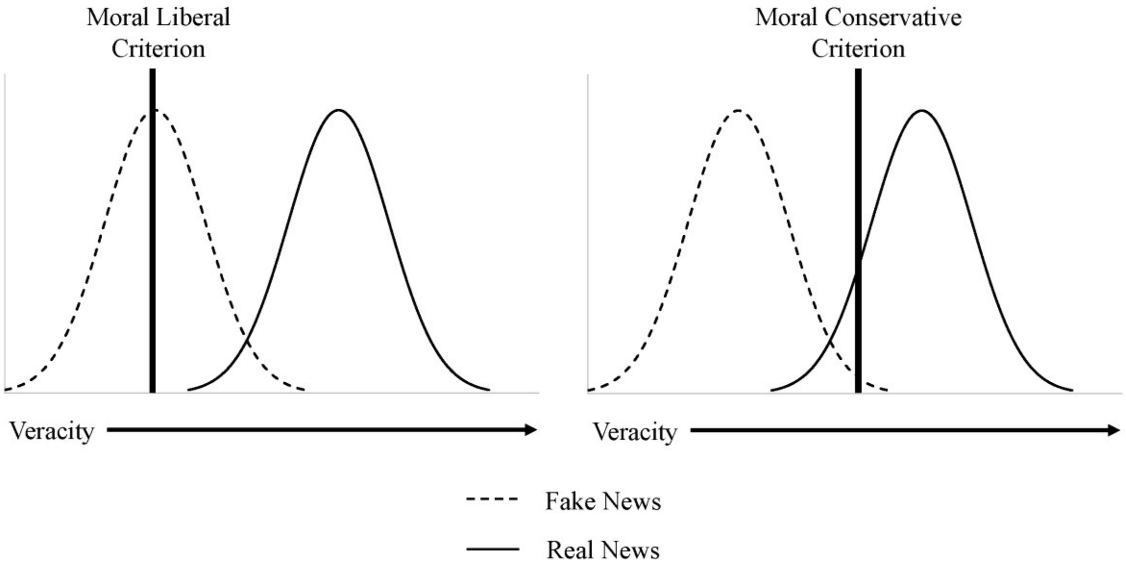


Figure 3. SDT *d′* scores reflecting accuracy in discriminating real news and fake news as a function of Cognitive Reflection Test scores and congruence of news content with participants' political orientation. Higher scores reflect greater accuracy in discriminating between real news and fake news. Reanalysis of data from Pennycook and Rand (2019).

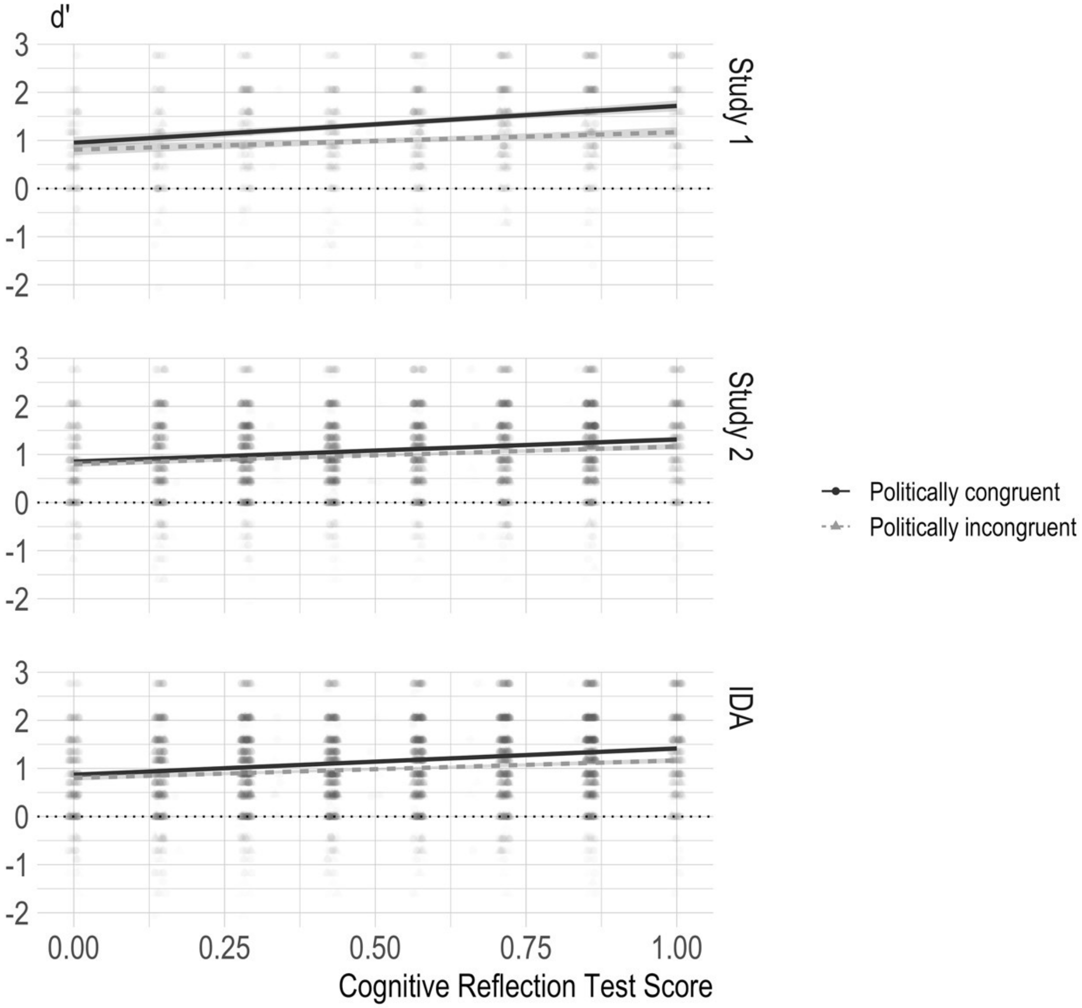


Figure 4. SDT *c* scores reflecting response bias in judging news headlines as real versus fake regardless of their veracity as a function of Cognitive Reflection Test scores and congruence of news content with participants' political orientation. Scores greater than zero reflect a response bias to judge news headlines as fake regardless of their veracity; scores lower than zero reflect a response bias to judge news headlines as real regardless of their veracity. Reanalysis of data from Pennycook and Rand (2019).

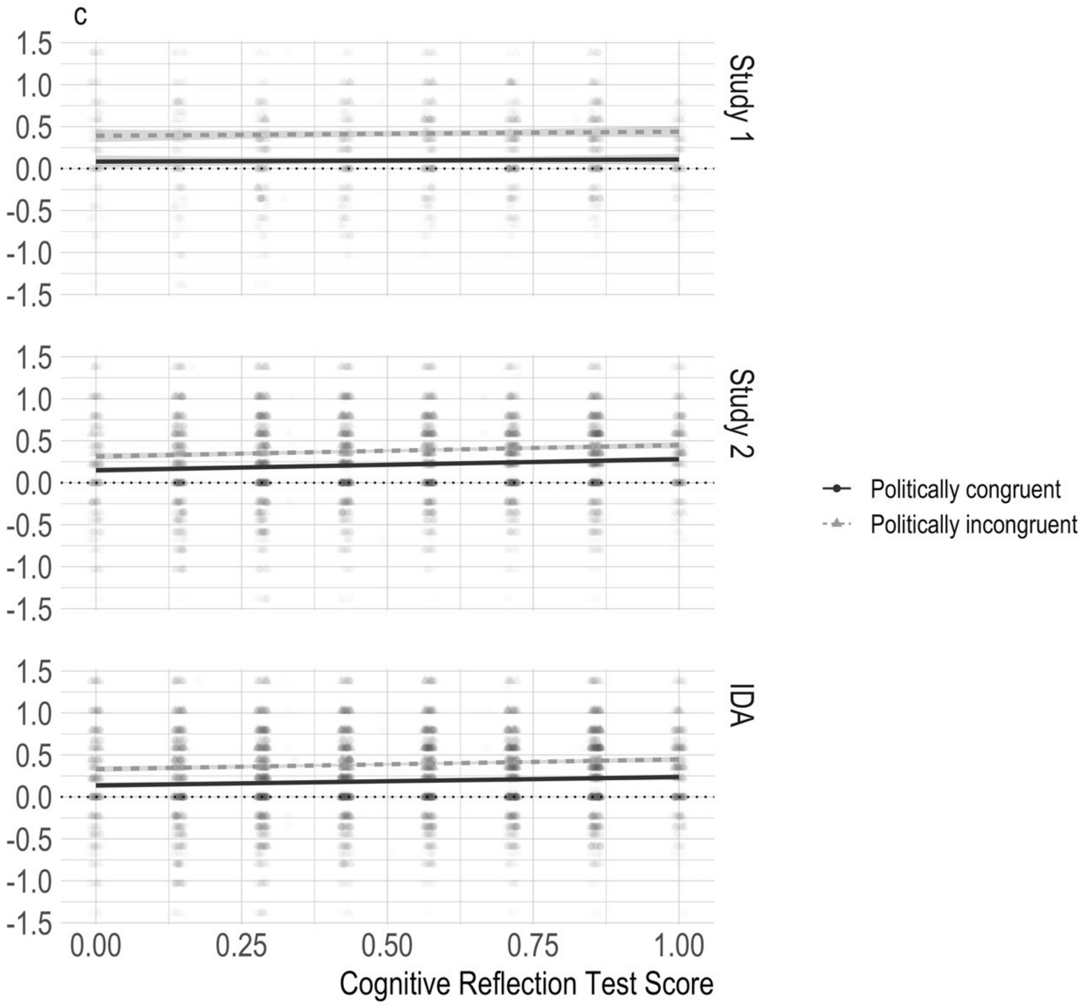


Figure 5. SDT *d′* scores reflecting accuracy in discriminating real news and fake news as a function of prior exposures. Higher scores reflect greater accuracy in discriminating between real news and fake news. Reanalysis of data from Pennycook, Cannon, and Rand (2018).

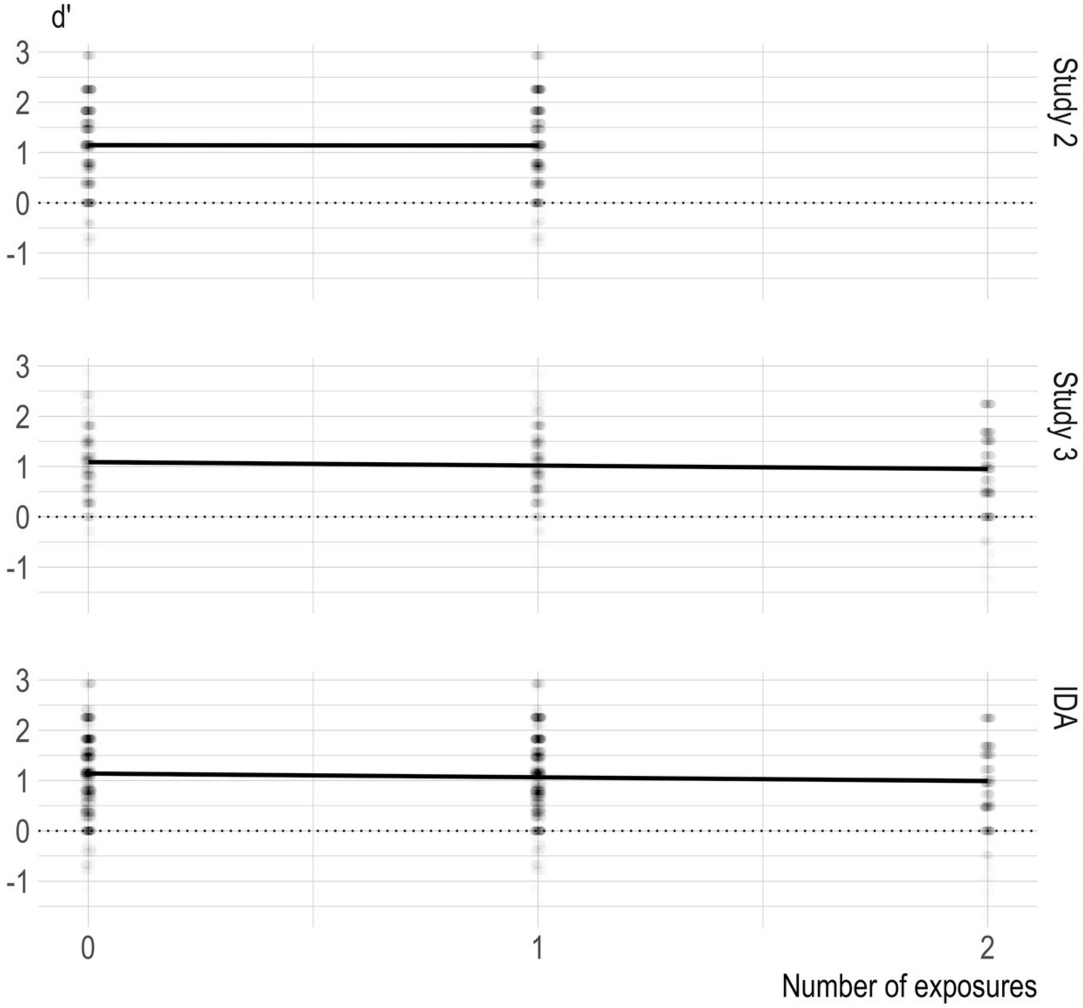
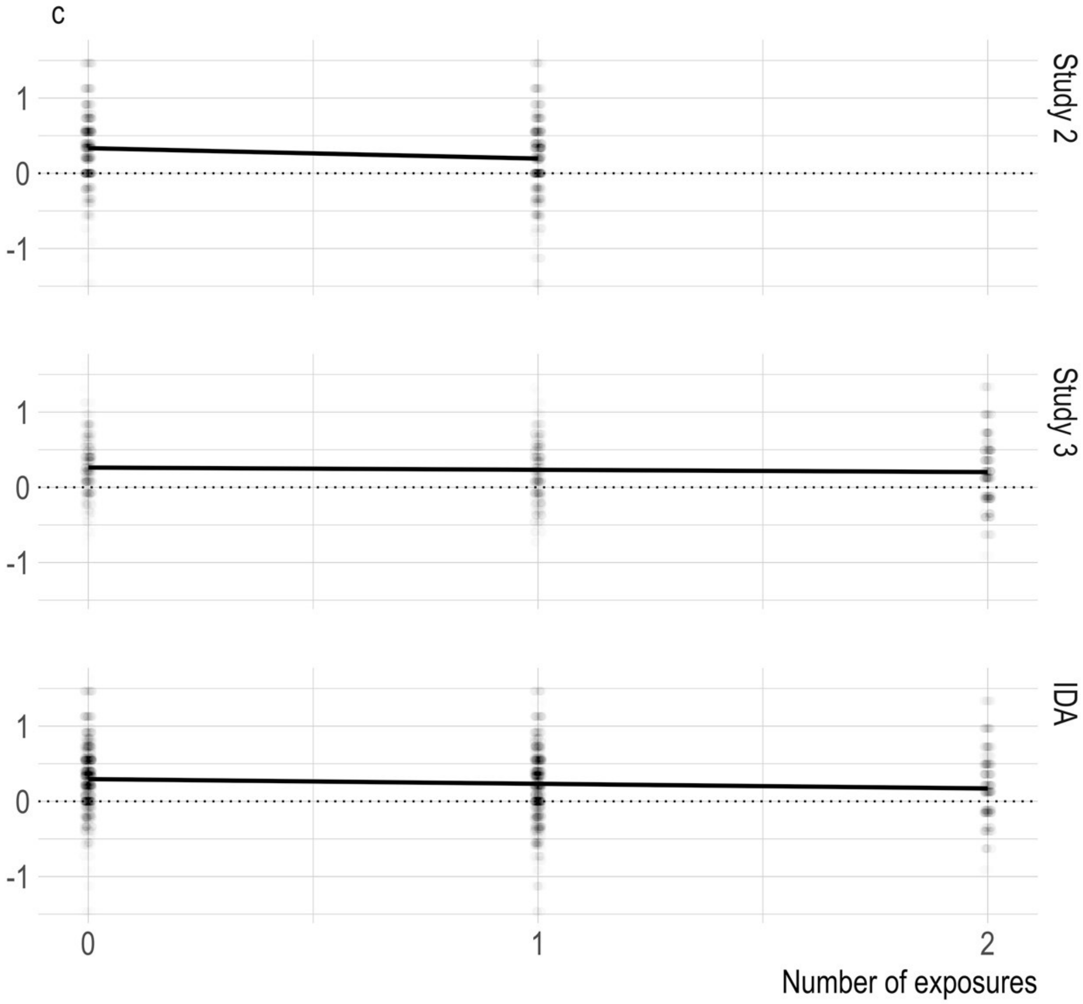


Figure 6. SDT *c* scores reflecting response bias in judging news headlines as real versus fake regardless of their veracity as a function of prior exposures. Scores greater than zero reflect a response bias to judge news headlines as fake regardless of their veracity; scores lower than zero reflect a response bias to judge news headlines as real regardless of their veracity. Reanalysis of data from Pennycook, Cannon, and Rand (2018).


c
Study 2
Study 3
IDA
Number of exposures