

Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking

Jon Roozenbeek* Rakoen Maertens† Stefan M. Herzog‡ Michael Geers§
Ralf Kurvers¶ Mubashir Sultan|| Sander van der Linden**

Abstract

Misinformation presents a significant societal problem. To measure individuals' susceptibility to misinformation and study its predictors, researchers have used a broad variety of ad-hoc item sets, scales, question framings, and response modes. Because of this variety, it remains unknown whether results from different studies can be compared (e.g., in meta-analyses). In this preregistered study (US sample; $N = 2,622$), we compare five commonly used question framings (eliciting perceived headline accuracy, manipulativeness, reliability, trustworthiness, and whether a headline is real or fake) and three response modes (binary, 6-point and 7-point scales), using the psychometrically validated Misinformation Susceptibility Test (MIST). We test 1) whether different question framings and response modes yield similar responses for the same item set, 2) whether people's confidence in their primary judgments is affected by question framings and response modes, and 3) which key psychological factors (myside bias, political partisanship, cognitive reflection, and numeracy skills) best predict misinformation susceptibility across assessment methods. Different response modes and question framings yield similar (but not identical) responses for both primary ratings and confidence judgments. We also find a similar nomological net across conditions, suggesting cross-study comparability. Finally, myside bias and political conservatism were strongly positively correlated with misinformation susceptibility, whereas numeracy skills and especially cognitive reflection were less important (although we note potential ceiling effects for numeracy). We thus find more support for an "integrative" account than a "classical reasoning" account of misinformation belief.

Keywords: misinformation, fake news, susceptibility, confidence, partisanship, active open-minded thinking, analytical thinking

*Department of Psychology, University of Cambridge. <http://orcid.org/0000-0002-8150-9305>

†Department of Psychology, University of Cambridge. <https://orcid.org/0000-0001-8507-5359>

‡Center for Adaptive Rationality, Max Planck Institute for Human Development. <https://orcid.org/0000-0003-2329-6433>

§Center for Adaptive Rationality, Max Planck Institute for Human Development. Department of Psychology, Humboldt University of Berlin. <https://orcid.org/0000-0003-0602-5893>

¶Center for Adaptive Rationality, Max Planck Institute for Human Development. <https://orcid.org/0000-0002-3460-0392>

||Center for Adaptive Rationality, Max Planck Institute for Human Development. Department of Psychology, Humboldt University of Berlin. <https://orcid.org/0000-0002-4499-7143>

**Department of Psychology, University of Cambridge. <https://orcid.org/0000-0002-0269-1744>

1 Introduction

The study of misinformation has garnered significant attention within the social and behavioural sciences (Van Bavel et al., 2020; van der Linden et al., 2021). A large variety of assessment tools have been developed to measure misinformation susceptibility (Loomba et al., 2021; Maertens, Götz, et al., 2022), investigate predictors of why people fall for misinformation (Pennycook & Rand, 2019; 2020; Van Bavel et al., 2021; Roozenbeek et al., 2020), and test the efficacy of interventions (Guess et al., 2020; Pennycook et al., 2021; Roozenbeek & van der Linden, 2019). In doing so, researchers have used a variety of question framings (e.g., eliciting the perceived reliability, manipulateness, trustworthiness, or accuracy of a set of items, usually news headlines or social media posts) and response modes (i.e., the number of response options, e.g., binary classification, 6-point, or 7-point rating scales). For instance, the work by Pennycook, Rand, and colleagues typically uses a set of real and false news headlines in a Facebook format, where participants are asked to rate the accuracy of each headline on a binary (e.g., “To the best of your knowledge, is this headline accurate? Yes/No”), 4-point, or 6-point scale (Pennycook et al., 2020, 2021; Pennycook & Rand, 2019; 2020). Similar framings and scales have been used by Fazio (2020) and Guess et al. (2020). Van der Linden, Roozenbeek and colleagues, on the other hand, tend to use social media posts (from Twitter or Facebook) or WhatsApp conversations as stimuli, either with or without source information, asking participants to rate the reliability (Basol et al., 2020; Maertens et al., 2021; Roozenbeek, Maertens, et al., 2021) or manipulateness (Basol et al., 2021; Saleh et al., 2021) of these posts on a 7-point scale. Other commonly used question framings in misinformation research include asking participants to rate items’ trustworthiness (McGrew, 2020; Roozenbeek, van der Linden, et al., 2022), credibility (Pehlivanoglu, 2021), and whether an item is real/true or fake/false (Maertens, Götz, et al., 2022; Swire et al., 2017).

In general, previous research has found that varying the question framings or response modes can have a significant impact on participants’ responses in a wide array of different domains. Bradburn (1982) and Schwartz (1999), for example, found that question wording matters a great deal when designing surveys (for a review, see Bruine de Bruin, 2011). Andrews (1984) showed that the number of answer scale categories had a big impact on data quality, indicating that the number of response options used in a survey could have a significant effect on the interpretation of different findings. Similarly, Preston and Colman (2000) and Revilla et al. (2014) also found that significant differences arise when varying response modes (see DeCastellarnau, 2018, for an overview). Within the context of misinformation research, this variability can have important consequences. For example,

This work was funded by the ESRC (ES/V011960/1), Cambridge Trust, IRIS Coalition (UK government, SCH-00001-3391), and JITSUVAX (EU Horizon 2020, 964728). M.G. is a fellow of the International Max Planck Research School on the Life Course (LIFE), Berlin. We thank Eward Cokely and Jinhyo Cho for comments.

Copyright: © 2022. The authors license this article under the terms of the [Creative Commons Attribution 3.0 License](https://creativecommons.org/licenses/by/3.0/).

Smith (1995) discusses how self-reported levels of Holocaust denial can vary depending on how a survey question is phrased that seeks to elicit the degree of knowledge that people have about the Holocaust.

No research to date has directly compared the response patterns that are produced when using different question framings and response modes to assess misinformation susceptibility. Hence, it remains unknown whether different studies that ostensibly seek to assess the same construct indeed do so. This is important, because if assessing misinformation susceptibility is robust to different question framings and response modes, then the results of such diverse studies will be directly comparable, for example in meta-analyses. If this is not the case, then not only are the outcomes of studies using different question framings and response modes not directly comparable, but a careful rethinking of which question framings and response modes tap into which exact construct will also be required. We address these key open questions in this study.

This study has two additional goals. First, we investigate the role of confidence judgments in the assessment of misinformation susceptibility across question framings and response modes. The accuracy of people's confidence in detecting misinformation is crucial for three reasons. Firstly, confidence influences whether people act on their initial (truth) judgment or seek additional information (Berner & Graber, 2008; Meyer et al., 2013), thereby making accurate confidence judgments a prerequisite for realising the need to verify information (Salovich & Rapp, 2020). Secondly, the level of confidence a person has in their beliefs affects their willingness and ability to defend these beliefs (Tormala & Petty, 2004). Individuals who are justifiably confident in their ability to assess the veracity of news content are thus less likely to fall for misinformation (Basol et al., 2020; Compton & Pfau, 2005). Thirdly, people listen more to more confident voices, especially in the absence of cues indicating competence (Price & Stone, 2003; Tenney et al., 2008), making it crucial to understand how well confidence signals competence. Little is currently known about the extent to which confidence assessments are influenced by the use of different question framings and response modes, leaving a knowledge gap in the cross-study comparability of confidence measures (Basol et al., 2020). Additionally, the relationship between various item ratings (e.g., reliability) and accompanying confidence judgments are unclear: Are primary ratings of news headlines an indication of the extent to which participants think the headlines possess a particular continuous property (e.g., more or less reliability), or are ratings simply an indication of the confidence with which they classified the headline into one of two categories (e.g., as being reliable vs. unreliable)? If the latter is the case, then the rating responses collected in previous research may need to be re-interpreted and could possibly be treated as a proxy for confidence (such as in re-analyses).

Second, there is an ongoing discussion about the predictors of misinformation susceptibility (Pennycook & Rand, 2021; Van Bavel et al., 2021). Such studies are (again) often conducted using psychometrically non-validated measures. To test various accounts of misinformation belief against each other, their predictions should be compared within

a common measurement framework. We therefore compare two overlapping accounts of misinformation susceptibility. The “classical reasoning” account of misinformation belief (Pennycook & Rand, 2019; 2020) argues that a lack of “reflexive open-mindedness” underlies belief in false news (Pennycook & Rand, 2020, p. 187), and that motivated or identity-protective thinking plays a relatively minor role (Pennycook & Rand, 2019, p. 48). This account emphasizes the role of analytical thinking in susceptibility to misinformation (Pennycook & Rand, 2020). Conversely, an “integrative account”, called for by Van Bavel et al. (2021) and van der Linden (2022), proposes that in addition to purely cognitive factors such as analytical skills, identity-protective thinking, “myside bias”, and political ideology are central factors in predicting misinformation susceptibility. Comparing these two accounts across different assessment methods could bring new insights not only into the nature of misinformation belief, but also into whether different assessment methods yield a comparable nomological net (i.e., a similar profile of predictor coefficients across different assessment methods of misinformation susceptibility).

We therefore explore how well four key psychological factors predict misinformation susceptibility across question framings and response modes: endorsement of actively open-minded thinking (AOT; Baron, 2019; e.g., Maertens, Götz, et al., 2022; Erlich et al., 2022); political ideology (Van Bavel & Pereira, 2018; Van Bavel et al., 2021); analytical thinking (as assessed by the cognitive reflection test or CRT; Pennycook & Rand, 2019; 2020); and numeracy skills (Roozenbeek et al., 2020). AOT is highly sensitive to acceptance of “myside bias” (Baron, 2019, p. 10; Svedholm-Häkkinen & Lindeman, 2018, p. 22), which “occurs when people evaluate evidence, generate evidence, and test hypotheses in a manner biased toward their own prior opinions and attitudes” (Stanovich et al., 2013), and is claimed to be one of the strongest psychological predictors of misinformation belief (Van Bavel et al., 2021, p. 96). Political ideology is a measure of partisanship, which is argued to predict false news detection ability (see Van Bavel et al., 2021; Gawronski, 2021). The CRT is commonly used as a proxy for analytical thinking, and is also argued to be a strong contributing factor to false news belief (Pennycook & Rand, 2019; 2020; 2021). Numeracy skills are also an indicator of analytical thinking ability, and were found to be the strongest predictor of lower belief in COVID-19 misinformation across 5 different countries (Roozenbeek et al., 2020). Although the “integrative account” acknowledges that analytical thinking plays a role in predicting misinformation susceptibility, it also emphasises the role of myside bias and partisanship. According to this model, we thus expect AOT and political ideology to be more consistent predictors of misinformation susceptibility than CRT and numeracy. Conversely, an extreme form of the “classical reasoning” account implies that CRT and numeracy skills are more robust predictors than AOT and political ideology.¹

¹We note that previous research has proposed that both open-mindedness (AOT) and analytical thinking (CRT) measure reflective reasoning ability, in accordance with the “classical reasoning” account (Erlich et al., 2022; Pennycook, Cheyne et al., 2020). However, we argue that there are important conceptual differences between the two scales: AOT is sensitive to “myside bias” (Baron, 2019, p. 10; Svedholm-Häkkinen & Lindeman, 2018, p. 22), whereas the CRT measures cognitive reflection (Frederick, 2005; Thomson &

1.1 The Misinformation Susceptibility Test (MIST)

To address the above questions, we make use of the Misinformation Susceptibility Test (MIST), a psychometrically validated scale that assesses misinformation susceptibility (Maertens, Götz, et al., 2022). The full version of the MIST (the MIST-20) consists of 10 real and 10 made-up (i.e., false) headlines (without source information and images, both of which can affect people's news evaluation, see Pehlivanoglu et al., 2021; Zillmann et al., 1999), obtained via a combination of factor analysis, classical test theory, and item-response theory models. These headlines were tested using Differential Item Functioning (DIF) analysis based on ideology (liberal–conservative), removing all items that lead to measurement inaccuracies due to their ideological slant. The 10 false headlines were created using GPT-2, a text-generating model developed by OpenAI that was trained on a large sample of false headlines. The 10 real headlines were taken from real and legitimate news sources. Because the psychometric properties of the test are known, the MIST is a strong instrument to evaluate misinformation susceptibility, and examine variations across question framings and response modes.

The MIST-8 is a shortened version of the MIST-20, consisting of the eight best-performing headlines from the MIST-20 (four false and four true; see Appendix). Although not preregistered, we report the results for both the MIST-20 and the MIST-8 throughout this paper, to illustrate minor variations that may arise when using subscales of larger item sets.

Performance on the MIST is scored according to three separate metrics: veracity discernment ability (**VDA**, the ability to discern true from false news), in addition to a real news score (**RNS**, accuracy in identifying real headlines) and fake news score (**FNS**, accuracy in detecting false headlines). These scores all correlate strongly with other item sets commonly used in misinformation research, such as the true and false COVID-19 headlines used by Pennycook et al. (2020). For a detailed discussion about the MIST's design, usage, and psychometric properties, see Maertens, Götz, et al. (2022). The Appendix lists the MIST-20 and MIST-8 headlines.

2 The present study

In this study, we use the MIST-20 and MIST-8 to compare five question framings (eliciting the *accuracy*, *manipulativeness*, *reliability* and *trustworthiness* of each MIST headline, and whether the headline is *real or fake*) and three response modes (6-point, 7-point and binary scales) commonly used in misinformation research, as well as the level of confidence

(Oppenheimer, 2016); within the context of misinformation susceptibility, we consider both scales to measure distinct constructs, although both could be broadly classified as analytical “reasoning”. See the Discussion section for additional discussion.

that people report to have in their judgment of each MIST headline. As preregistered², we expect the MIST to yield similar responses for primary headline ratings as well as confidence judgments across different question framings and response modes³.

In order to compare the “integrative” and “classical reasoning” accounts of misinformation belief and compare the nomological net of different assessment methods, we conduct a series of analyses (Pearson’s and disattenuated correlations, as well as linear regressions) to examine how MIST-20 and MIST-8 veracity discernment ability (VDA) correlates with actively open-minded thinking (AOT), political ideology, analytical thinking (CRT), and numeracy skills, across response modes and question framings.

Our Open Science Framework (OSF) page contains all the information required to replicate our methods and results, including the raw and cleaned datasets, Qualtrics survey, preregistration, supplementary tables and figures, and our analysis and visualisation scripts: <https://osf.io/b9m3k/>. Our preregistration can be accessed here: <https://aspredicted.org/7ht5z.pdf>.

3 Method

3.1 Sample and procedure

We conducted our study on Prolific Academic (Peer et al., 2017) using the survey software Qualtrics. We followed previously established guidelines on initial scale development, which recommend recruiting at least 300 respondents per condition (Maertens, Götz, et al., 2022; Clark & Watson, 2019; Boateng et al., 2018). As preregistered, we sought to recruit a United-States-based sample of 2,674 participants, with 334 participants per condition. After excluding participants who failed attention checks, we ended up with a slightly smaller final sample of $N = 2,622$, consisting of 50.6% women (46.9% men, 1.7% non-binary, 0.6% other, 0.3% prefer not to say) with a mean age of 37.1 ($SD = 13.7$). Participants were, on average, left-leaning (political ideology: $M = 3.07$, $SD = 1.74$, on a 7-point scale), and 66.3% reported having obtained at least a bachelor’s degree. Most participants lived in the South (34.4%) or West (25.0%) of the United States (with 22.2% reporting living in

²We preregistered the following null hypotheses: **H0a**: Different question framings for participants’ scores (measured on the same 7-point scale) on the 20-item Misinformation Susceptibility Test (MIST-20) measure the same latent construct; and **H0b**: Different response modes (7-point Likert scale, 6-point Likert scale, and binary scale) for participants’ scores on the MIST-20 measure the same latent construct.

³We deviate from our preregistration in the following ways: **1**) the preregistration mentions only the 20-item MIST (MIST-20), whereas we also present the results for the shorter 8-item MIST (MIST-8), which consists of a subset of the 8 best-performing MIST-20 headlines (for a detailed overview of the MIST-8 and its psychometric properties, see Maertens, Götz, et al., 2022). **2**) We preregistered that we would look at headline recognition (i.e., whether MIST scores are related with whether people indicate having seen a headline before). Due to space limitations, we will not do this here, but save this for a future publication. **3**) We preregistered that we would use lavaan’s *lavTestLRT* feature for invariance testing. We now additionally use lavaan’s *summary()* function to find the robust model fit measures, and provide each model’s Cronbach’s α and McDonald’s ω as additional reliability indicators.

the North-East and 18.4% in the Mid-West). Participants were paid GBP 0.55 for their participation. See Table S1 of the for the full sample composition.⁴

The procedure of this study was as follows: after providing informed consent, participants were randomly assigned to one of eight conditions, which differed in their combination of question framings and/or response modes (see below); these combinations were selected based on their use in previous research. In each condition, participants rated the MIST-20 headline set (headline presentation order was random for each participant). Each condition used a different question framing and/or response mode for the primary judgment:

1. **Accuracy (6 pt.) (n = 326):** “How *accurate* do you find this headline?”, 1 being “not at all” and 6 being “very” (Guess et al., 2020; Pennycook et al., 2020).
2. **Accuracy (7 pt.) (n = 336):** “How *accurate* do you find this headline?”, 1 being “not at all” and 7 being “very”.
3. **Manipulativeness (7 pt.) (n = 330):** “How *manipulative* do you find this headline?”, 1 being “not at all” and 7 being “very” (Basol, Roozenbeek, et al., 2021; Saleh, Roozenbeek, et al., 2021).
4. **Reliability (7 pt.) (n = 331):** “How *reliable* do you find this headline?”, 1 being “not at all” and 7 being “very” (Basol et al., 2020; Roozenbeek & van der Linden, 2019).
5. **Trustworthiness (7 pt.) (n = 330):** “How *trustworthy* do you find this headline?”, 1 being “not at all” and 7 being “very” (McGrew, 2020; Roozenbeek, van der Linden, et al., 2022).
6. **Real – Fake (6 pt.) (n = 315):** “This headline is...”, 1 being “*real*” and 6 being “*fake*”.
7. **Real – Fake (7 pt.) (n = 316):** “This headline is...”, 1 being “*real*” and 7 being “*fake*”.
8. **Real – Fake (Binary) (n = 338):** “This headline is...”, real or fake as a *binary* judgment (Maertens, Götz, et al., 2022).

Conditions 2, 3, 4, 5, and 7 differ in their question framing but all use a 7-point scale. Conditions 1 and 6 use different question framings but use a 6-point scale. Conditions 1 and 2 and conditions 6, 7, and 8 use the same question framings (accuracy and real-vs-fake, respectively), but different response modes (6- and 7-point scales in conditions 1 and 2, and 6-point, 7-point, and binary scales in conditions 6, 7, and 8). After indicating their judgment of a headline, participants were also asked to indicate their confidence in their judgment (“How confident are you in your judgment?”, 1 being “not at all” and 7 being “very”; e.g., Saleh et al., 2021).

⁴All references beginning with S refer to the [Supplement](#).

Participants then completed a series of demographic and other questions in the following order: age, gender, education level, political ideology (from 1 “very liberal” to 7 “very conservative”), political party identification (Democrat/Republican/Independent/Other), US geographic region (West/Mid-West/South/North-East), news consumption (how often people check the news, from 1 “never” to 5 “all the time”), social media use (from 1 “never” to 5 “all the time”), the 10-item actively open-minded thinking scale (AOT; for the specific scale used, see Baron et al., 2022), the 3-item cognitive reflection test (CRT-2, hereafter referred to as CRT; Thomson & Oppenheimer, 2016), the 3-item Schwartz numeracy test (Schwartz et al., 1997), and a single-item risk literacy test (“which represents the highest risk of something happening: 1 in 10 / 1 in 100 / 1 in 1000”; see Dryhurst et al., 2020; Roozenbeek et al., 2020). The Schwartz test and risk literacy test were combined into a single numeracy score, following Roozenbeek et al. (2020). We also recorded participants’ reaction times for both the primary and confidence ratings, as well as whether they had seen or heard about each MIST headline before.⁵ Finally, participants were debriefed about the nature and purpose of the study. Figure 1 shows the study design (with the headline recognition task excluded).

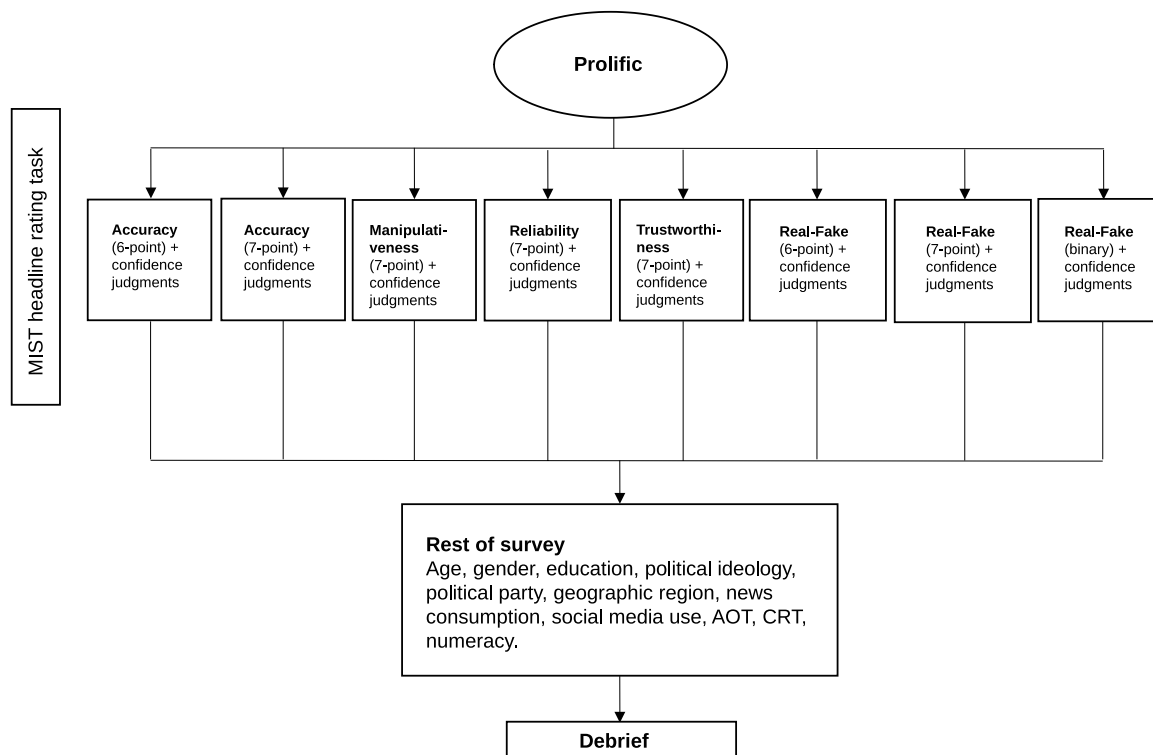


FIGURE 1: Flowchart showing the study design.

⁵We do not explore headline recognition in this paper.

3.2 Analyses

We preregistered and followed the following analysis plan: To determine whether different question framings and response modes measure the same latent construct, we used structural equation modelling (SEM). Specifically, we conducted a measurement invariance test (testing for three sequential levels of invariance: configural, metric, and scalar) in *lavaan* on the five question framings (only for conditions with 7-point scales) and three response modes (for the 6- and 7-point and binary scales for the real-vs-fake question framing, as well as for the 6- and 7-point scales for the accuracy question framing), for both the MIST-20 and the MIST-8. We also list here the model fit values for each group. Achieving configural invariance — the lowest level of invariance — means that the overall factor structure of the SEM exhibits a similar fit in each group, or that there is a “qualitatively invariant measurement pattern of latent constructs across groups” (Xu & Tracey, 2017, p. 75). A configural invariance test fits the model onto each group, while leaving factor loadings (the strength of each item’s — in our case a MIST headline — relation to the latent factor) and item intercepts (each item’s initial value) free to vary across groups. Metric invariance (the second level of invariance) is achieved when factor loadings (but not item intercepts) are equivalent across groups, indicating that each scale item (MIST headline) loads onto the model’s latent factor in a similar manner. Finally, scalar invariance means that both factor loadings and item intercepts are equivalent across different groups, indicating that there is very little difference in terms of scale properties between groups (Lee, 2018).

Because we expected the structure of misinformation susceptibility to be the same across question framings and response modes, more invariance is more evidence in favour of this hypothesis. However, we recognise that changes in response modes and question framings could result in small changes in the interpretation of individual items, reflected in the factor loadings and intercepts, while still maintaining the general factor structure of the MIST-20 and MIST-8. Scalar invariance would provide excellent evidence and metric invariance very good evidence; we treat *at least* configural invariance across groups as valid support for the hypotheses (although note that this definition was not preregistered). To test for configural invariance, we fit a multiple-group SEM model and looked at the model fit indices. We expected model fit indices of CFI/TLI > .90 and SRMR/RMSEA < .10 for good fit, and CFI/TLI > .95 and SRMR/RMSEA < .06 for excellent fit (Clark & Watson, 2019; Finch & West, 1997; Hu & Bentler, 1999; Pituch & Stevens, 2015; Schumacker et al., 2015). To test for metric and scalar invariance, we used a standard chi square invariance test.

Additionally (not preregistered), we compared the eight conditions (using ANOVAs) according to three metrics introduced in Maertens, Götz, et al. (2022), each assessing a different aspect of misinformation susceptibility: veracity discernment ability (i.e., accuracy in discerning real news from false news; **VDA**), real news score (i.e., accuracy in identifying real headlines; **RNS**), and fake news score (accuracy in identifying false headlines; **FNS**). VDA is calculated by standardising each of the responses on a scale from 0 (most incorrect) to 1 (most correct) and taking the mean of the item scores. For more information about how

the RNS and FNS scores are calculated, see Supplementary Analysis S1.⁶

With respect to the confidence ratings, we employ an exploratory approach rather than formalised statistical tests. We descriptively compare participants' mean confidence ratings across conditions. In addition, we investigate the association between the primary headline judgments and the confidence judgments by constructing an implied full-range confidence that ranges from very confident that an item is inaccurate (reliable/non-manipulative, etc.), to very confident that the item is accurate (unreliable/manipulative, etc.).⁷ Then we descriptively compare the within-participant Spearman correlations between a participant's primary and full-range confidence judgments across conditions; these analyses do not include the binary real–fake condition because in that condition only two distinct responses are possible (“real” or “fake”) and thus no continuous associations can be investigated.

To test whether actively open-minded thinking (AOT), political ideology, analytical thinking (CRT), or numeracy skills are most consistently associated with misinformation susceptibility, we compute both standard and disattenuated Pearson's correlations between MIST-20 and MIST-8 veracity discernment ability (VDA) and AOT, political ideology, CRT, and numeracy. In addition, we report the correlations of each of these variables with news consumption and participants' reaction time for the MIST headline ratings (log-transformed); we include these variables to check whether reaction time and news consumption may serve as confounds for the four variables mentioned above. We also estimate a series of linear regressions with AOT, CRT, numeracy and political ideology simultaneously predicting veracity discernment ability as well as participants' real and fake news scores (RNS and FNS).

4 Results

4.1 Question framings and response modes

We performed a measurement invariance test on the five question framings using a 7-point scale (*accuracy, manipulateness, reliability, trustworthiness*, and the 7-point *real-vs-fake* scale) and the two sets of response modes (the two accuracy conditions and the three real-vs-fake conditions). Table S3 shows the fit values for the configural invariance models across all comparisons, for both the MIST-20 and MIST-8.

⁶Although we do not explore this in the present paper, we also computed the area under the receiver operator curve (AUC) for each condition, using the trapezoid rule. We find that veracity discernment ability (VDA) and the AUC are highly similar in each condition, with all Spearman correlations > 0.86 and all Pearson correlations > 0.89 (interestingly, AUC and VDA are mathematically the same in the real – fake binary condition, $r_{Spearman} = 1.00$ and $r_{Pearson} = 1.00$). See Figure S6 and Table S24 for detailed results.

⁷For this full-range confidence we binarised the primary judgments in all conditions. For the conditions with a 7-point rating scale, the middle-most response (“4”) does not imply one or the other decision; here we randomly assigned the judgment to one of the two binary categories. Furthermore, both the primary and full-range confidence judgments are aligned so that higher values point to “good” ratings (i.e., real, trustworthy, non-manipulative, etc.).

With respect to question framings, we found no configural invariance for the MIST-20, indicating that question framings change the psychometric properties of the MIST-20 substantially. For the MIST-8, we found configural invariance but no metric invariance ($\Delta\chi^2 = 73.83, p < .001$). These results provide partial support for the hypothesis that different question framings measure the same latent construct.

With respect to response modes, we likewise found no configural invariance for the MIST-20. For the MIST-8, we found metric measurement invariance across all response modes for all three real-vs-fake conditions ($\Delta\chi^2 = 18.51, p = .101$) as well as for the two accuracy conditions ($\Delta\chi^2 = 3.42, p = .755$). We also find scalar invariance for the 6- and 7-point real-vs-fake scales ($\Delta\chi^2 = 7.11, p = .210$). These results indicate that using different response modes does not alter the psychometric properties of the MIST-8 but does alter the properties of the MIST-20, providing partial support for the hypothesis that different response modes measure the same latent construct.

Furthermore, looking at the fit values for the eight conditions (see Table S3), we see that for six out of eight conditions, the MIST-8 SEMs had good fit values, further demonstrating internal consistency across models. Only one of the MIST-20 models (the binary real-vs-fake condition) showed a good fit. However, the MIST-20 generally has a good reliability (McDonald's $\omega > .70$; McDonald, 1999) in all eight conditions, indicating that the MIST-20 still provides a reliable measure of misinformation susceptibility across all response modes and question framings. Overall, these results show that, although varying question framings and response modes does result in variations in response patterns (particularly for the MIST-20), these variations are relatively minor.

We also compared (not preregistered) the eight conditions in order to gain more insight into between-condition variability in MIST veracity discernment ability (VDA), real news score (RNS), and fake news score (FNS). Figure 2 shows that all three scores are comparable across conditions, except for the binary real-vs-fake scale, which is significantly different from all other conditions in that participants in this condition have higher VDA and RNS (but not FNS). Overall, these results are in line with the results from the SEM analysis, and further support the notion that, minor variations notwithstanding, participants' MIST headline ratings are similar across question framings and response modes. See Table S4 for the descriptive statistics and Tables S5-S10 for the Games-Howell post-hoc tests for each of the three measures.

4.2 Confidence judgments

Figure 3 shows the distribution of confidence ratings per condition. We find that confidence distributions follow a comparable pattern across all conditions, except for the manipulativeness condition, in which participants gave higher confidence scores (see the non-overlapping confidence intervals in Figure 3). Furthermore, invariance tests using SEMs suggest that configural invariance (but not metric or scalar invariance) was achieved across all conditions (see Table S11). Thus, with the exception of the “manipulativeness” question framing

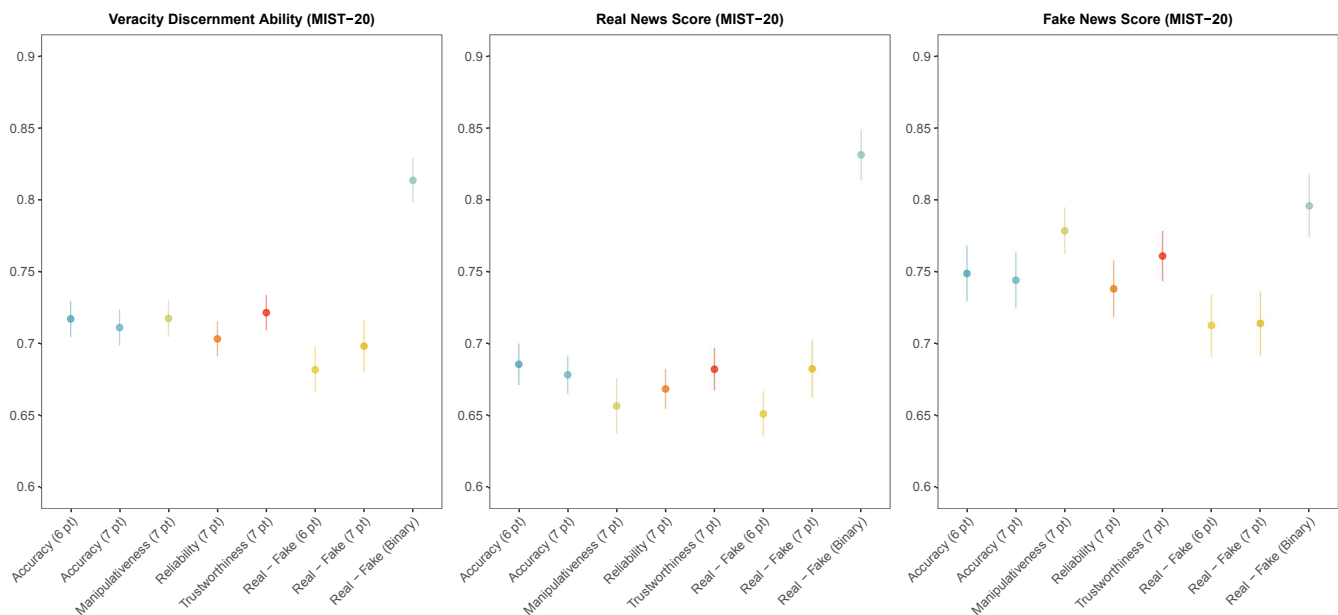


FIGURE 2: Point-range plots for MIST-20 veracity discernment ability, real news score, and fake news score, by condition. Dots represent the means, vertical lines represent the 95% confidence interval. See Figure S1 for the corresponding MIST-8 figure and Table S4 for the descriptive statistics.

condition, the results support the notion that confidence judgments of real news and false news are not meaningfully affected by the use of different question framings or response modes.

Figures S2 and S3 further show the relationship between participants' primary judgments and their confidence judgments across conditions, again confirming that both measures are very similar. Figure S4 shows the within-participant Spearman correlations between the primary (headline) and confidence ratings. This correlation is substantial in all conditions (all group medians $> .9$). These results largely support the notion that MIST headline ratings and confidence judgments measure the same latent construct.

4.3 Comparing two accounts of misinformation susceptibility

To test the “classical reasoning” account against the “integrative” account of misinformation belief, we preregistered exploratory analyses for actively open-minded thinking (AOT), the cognitive reflection test (CRT), and numeracy scales, as well as a single-item measure of political ideology. Table 1 shows the standard and disattenuated Pearson's correlations between veracity discernment ability (VDA), AOT, CRT performance, numeracy skills, and political ideology, news consumption, and reaction time for MIST headline ratings (log-transformed), separately for the MIST-20 and the MIST-8. The table displays the results for all eight conditions pooled together; see Table S12 for the results per condition. Figure

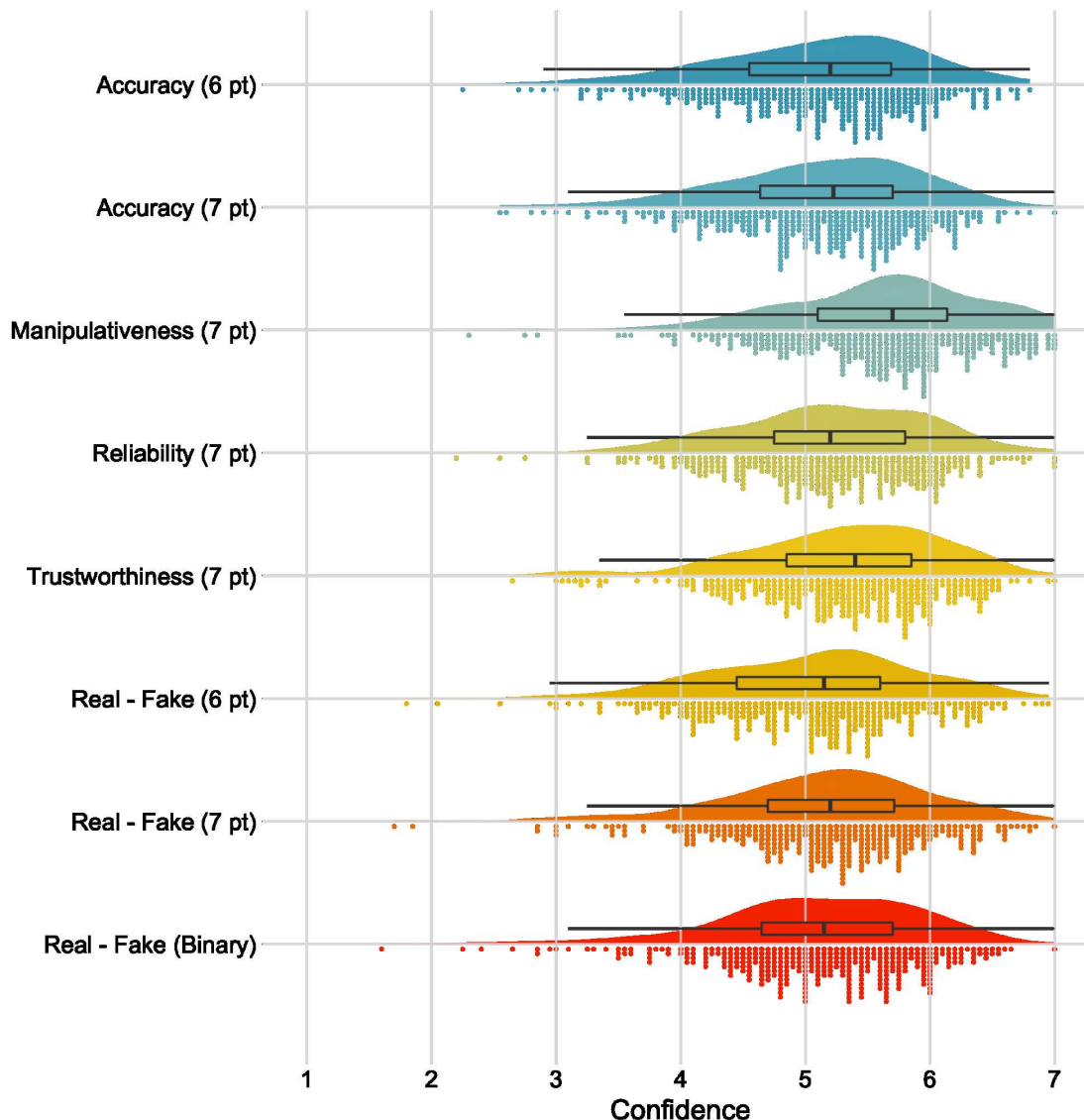


FIGURE 3: MIST-20 confidence ratings (1 being “not at all confident” and 7 being “very confident”) per condition, irrespective of the accuracy of the primary judgments. Per condition, the distribution is summarised by a boxplot (not showing outliers), a point range (showing the median and its 95% percentile-bootstrapped confidence interval), density plot, and a dot plot. The width of a boxplot is proportional to the square root of the number of participants in the respective distribution.

4 shows the correlations between VDA and AOT, CRT, numeracy, and political ideology (respectively) separated by condition in a series of scatterplots with LOESS curves.

AOT was most strongly correlated with VDA (i.e., lower misinformation susceptibility), followed by political ideology (with participants identifying as left-wing showing generally higher veracity discernment), before numeracy skills and finally CRT performance. Neither news consumption nor reaction time are strongly correlated with VDA. A series of z -tests

TABLE 1: Pearson’s correlations (green), Cronbach’s alpha (blue), and disattenuated correlations (yellow) between Veracity Discernment Ability (VDA), actively open-minded thinking (AOT), cognitive reflection test performance (CRT), numeracy test performance, political ideology (1-7, 1 being “very liberal” and 7 being “very conservative”), news consumption, reaction time to MIST veracity judgments (log-transformed), and confidence in these judgments. The table shows the results for both the MIST-20 and the MIST-8, for all 8 conditions pooled together. Significant Pearson’s correlations at $p < 0.05$ are marked in bold. See Table S14 for the z-tests comparing the correlation coefficients. See Tables S12 and S13 for the correlations and z-tests separated by condition, which show highly similar patterns. See also Table S25 for the correlations for Democrats and Republicans separately.

MIST-20	VDA	AOT	CRT	Numer.	Politics	News	RT	Conf.
VDA	0.80	0.54	0.22	0.42	-0.37	0.16	-0.11	0.37
AOT	0.43	0.78	0.24	0.38	-0.32	0.10	-0.02	0.29
CRT	0.15	0.16	0.58	0.57	-0.08	-0.07	-0.05	0.01
Numeracy	0.25	0.23	0.30	0.46	-0.17	-0.05	-0.12	0.18
Political Ideol.	-0.34	-0.28	-0.06	-0.11	-	-0.09	0.04	-0.14
News consump.	0.14	0.08	-0.06	-0.03	-0.09	-	0.00	0.27
Reaction time	-0.08	-0.01	-0.03	-0.07	0.04	0.00	0.66	0.01
Confidence	0.31	0.24	0.01	0.11	-0.13	0.26	0.01	0.87
MIST-8	VDA	AOT	CRT	Numer.	Politics	News	RT	Conf.
VDA	0.61	0.57	0.24	0.43	-0.39	0.19	-0.11	0.44
AOT	0.39	0.78	0.24	0.38	-0.32	0.10	-0.03	0.30
CRT	0.15	0.16	0.58	0.57	-0.08	-0.07	-0.07	0.01
Numeracy	0.23	0.23	0.30	0.46	-0.17	-0.05	-0.06	0.18
Political ideol.	-0.31	-0.28	-0.06	-0.11	-	-0.09	0.03	-0.13
News consump.	0.15	0.08	-0.06	-0.03	-0.09	-	0.01	0.29
Reaction time	-0.06	-0.01	-0.03	-0.02	0.02	0.00	0.41	-0.04
Confidence	0.29	0.23	0.01	0.10	-0.11	0.24	-0.02	0.72

comparing correlation coefficients (see Table S14) further shows that the raw correlation between VDA and AOT is significantly stronger than for all other variables (all p -values < 0.001). In addition, political ideology is more strongly correlated with VDA than both numeracy and CRT (all p -values < 0.001). However, note that the disattenuated correlation for numeracy is closer to that for AOT, and many subjects were at ceiling on the numeracy measure itself, thus casting doubt on the sufficiency of the correction for its unreliability.

Separating the data by condition shows a similar pattern: AOT and political ideology are strongly and consistently correlated with VDA, whereas CRT and numeracy show weak

or no correlations. The only exceptions to this pattern are the binary real-vs-fake condition, where none of the correlations between VDA and the four other variables differ significantly from one another, and the 7-point real-vs-fake condition, where the correlations between VDA and political ideology, numeracy and CRT are not significantly different; see Tables S12 and S13.

To assess the unique predictive contributions of each covariate, we estimated a series of linear regressions with veracity discernment ability (VDA) as the dependent variable and AOT, CRT, numeracy, and political ideology as the independent variables, both for the MIST-20 and the MIST-8.⁸ These regression models corroborate the above findings based on zero-order correlations: AOT is the strongest predictor of higher veracity discernment ability in all conditions, followed by political ideology, numeracy skills, and lastly CRT performance (which does not significantly predict veracity discernment in any conditions except the binary real-vs-fake condition; see Tables S17 and S18). Parallel regression models for the real and fake news scores (RNS and FNS) further corroborate these results (see Tables S19 and S20). Finally, to examine whether the partisan slant of the MIST headlines might influence these findings (particularly because political ideology is such a strong predictor of veracity discernment), we ran the same regression models as above but with the three most partisan (i.e., right-leaning) false MIST headlines excluded; doing so does not alter our conclusions (see Table S22).⁹ Finally, the nomological nets for the MIST-20 and MIST-8 are highly similar across all conditions (see Tables S15-S20), buttressing the earlier finding that different question framings and response modes are broadly comparable when measuring misinformation susceptibility.

Overall, we thus found more support for the “integrative” than the “classical reasoning” account of misinformation belief: actively open-minded thinking (i.e., “myside bias”) and political ideology are both robust predictors of misinformation susceptibility, whereas classical analytic thinking (as measured by both CRT performance and numeracy skills) is not.

5 General Discussion

Misinformation susceptibility has become a popular topic of academic research in recent years. To assess how susceptible individuals are to misinformation, researchers have

⁸A Durbin-Watson test shows that the residuals are uncorrelated ($d = 2.03$, $p = 0.412$, autocorrelation = -0.018). The diagnostics plots further show that the models’ normality assumption is met, see Figure S5. To check for potential multicollinearity, we calculated the variance inflation factors (VIFs) for AOT, CRT, numeracy, and political ideology separately for each condition (all VIFs < 1.273 , indicating no meaningful multicollinearity; see Table S23). Including other variables in the regression models such as news consumption, reaction time and confidence scores does not meaningfully change the results reported here; see Tables S15 and S16.

⁹As an extra collinearity check, we ran the same regression model but with AOT excluded from the model. When doing so, political ideology becomes the strongest predictor of veracity discernment, before numeracy and CRT, further indicating that there is no collinearity between AOT and CRT (see Table S21).

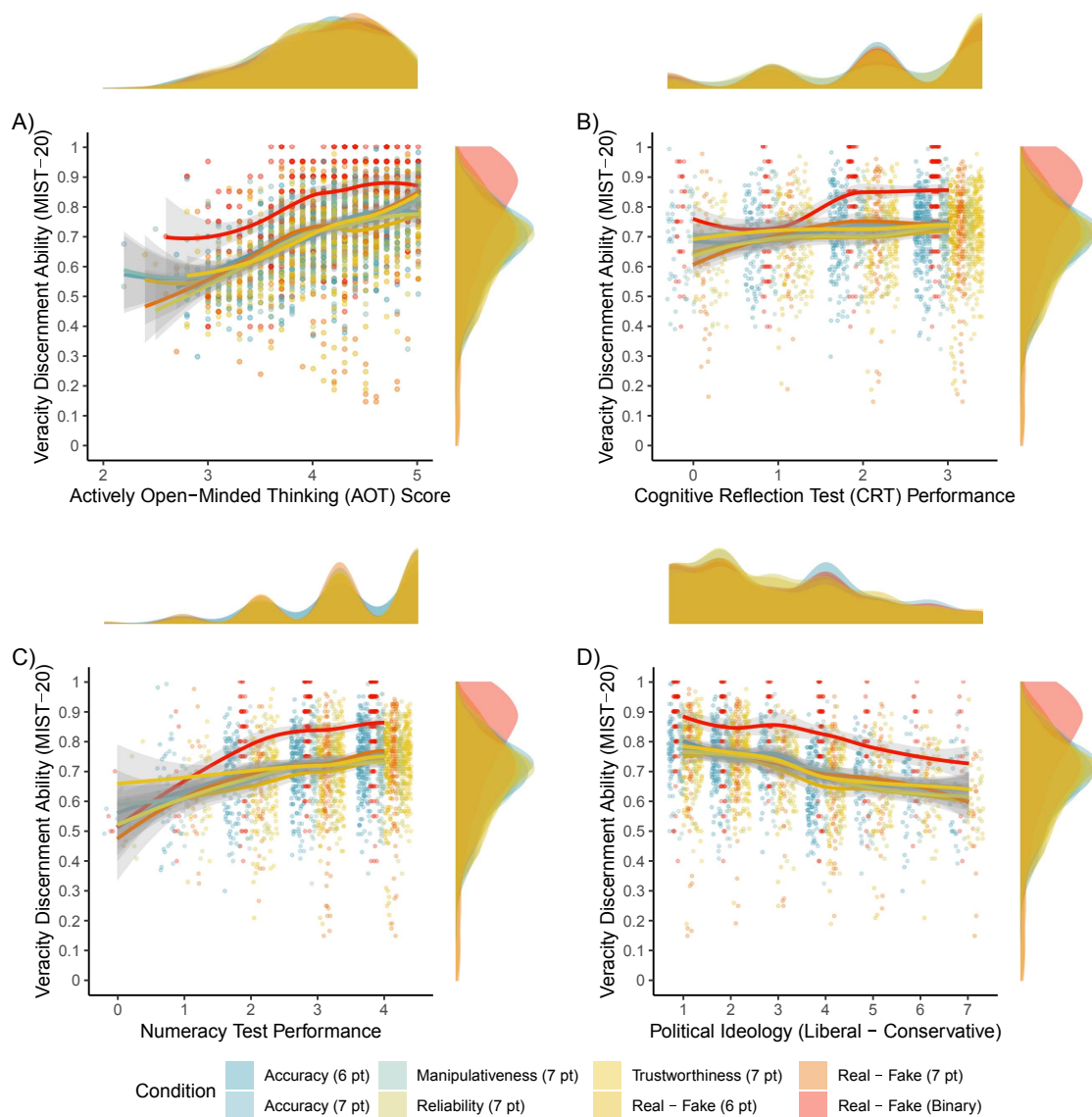


FIGURE 4: Actively Open-Minded Thinking (AOT; top left), Cognitive Reflection Test performance (CRT; top right), numeracy test performance (bottom left) and political ideology (liberal–conservative, bottom right), set against MIST-20 veracity discernment ability (VDA), by condition. Curves and confidence bands show robust LOESS curves (locally estimated scatterplot smoothing using re-descending M estimator with Tukey’s biweight function) and their 95% confidence bands.

used a variety of (often ad-hoc) measures, scales, and test item sets, as well as different question framings. While yielding impressive insights, this research has suffered from a lack of standardisation, and thus unclear cross-study comparability. To address this, we set out to examine whether measuring misinformation susceptibility is robust across different question framings and response modes. Moreover, we tested whether confidence judgments are affected by the use of different question framings and response modes, and

whether confidence judgments measure the same construct as the primary misinformation ratings. Finally, we tested two well-known accounts of misinformation susceptibility against each other across different assessment methods, using a psychometrically validated scale: the “integrative” account (Van Bavel et al., 2021) and the “classical reasoning” account (Pennycook & Rand, 2019; 2020).

5.1 Question framings and response modes

While there are differences across different response modes and question framings when measuring misinformation susceptibility, these differences appear to be minor, particularly for the MIST-8. A confirmatory factor analysis of different question framings (all using the same 7-point scale) showed that at least configural invariance was achieved across conditions for the MIST-8, indicating a (qualitatively) invariant pattern of measurement of latent constructs across conditions (Lee, 2018). Thus, while using different question framings for the MIST-8 does not result in *exactly* the same response patterns, they are similar enough to be broadly comparable. The results are even more robust for the different response modes, showing metric, and in some cases even scalar invariance across conditions, indicating that binary, 6-point, and 7-point scales can be expected to yield highly similar results, at times even down to the level of item intercepts.

For the MIST-20, the results are less clear: Although the fit measures of the SEMs are close to achieving configural invariance, they do not quite do so (see Table S3). However, a supplementary similarity test (looking separately at participants’ veracity discernment ability, real news score, and fake news score; Maertens, Götz, et al., 2022) showed that the between-condition variability for both question framings and response modes, although it does exist, is small (see Figure 2). These results offer external support for the idea that studies within misinformation research that are conceptually about the same thing (e.g., testing the efficacy of an anti-misinformation intervention using a set of test items) can be meaningfully compared to one another, for example in a meta-analysis.¹⁰

The observed differences between the MIST-20 and the MIST-8 might be a function of the quality of the scale that is used to measure misinformation susceptibility. For example, the MIST-20 uses a wider range of test items than the MIST-8, and therefore potentially measures misinformation susceptibility with higher precision and more reliability (see Table S3). However, the MIST-8 uses only the best and most predictive items of the MIST-20, usually resulting in a much better model fit (again see Table S3). Our findings therefore indicate that better items result in better stability across different response modes and question framings. As most studies that measure misinformation susceptibility use ad-hoc scales and/or tests of limited quality, the measurements may thus differ across response

¹⁰Note that our analysis of the robustness of the MIST’s nomological net (i.e., the pattern with which covariates predict the dependent variable) showed that the nomological net is highly robust across question framings and response modes, in further support of our finding that there are no major variations in expected responses when question framing and/or response mode are manipulated.

modes and questions framings, highlighting the importance of using psychometrically validated tests.

5.2 Confidence judgments

With respect to confidence, we find that primary headline ratings (e.g., judging the accuracy, reliability, or trustworthiness of an item) and confidence judgments (i.e., the confidence in one's primary judgment of the item) largely measure the same latent construct. Irrespective of question framing, we find very strong and similar associations between confidence judgments and primary (headline) judgments. Furthermore, we find support for the notion that confidence judgments are largely unaffected by the use of different response modes and question framings. That is, the average level of confidence is comparable across conditions.

The “manipulativeness” question framing (i.e., “how *manipulative* do you find this headline?”) behaved somewhat differently than the other (7-point scale) question framings, despite its acceptable fit values in the SEM (see Table S3). The mean confidence judgments were higher in this category compared to all other categories (see Figure 3). It is possible that these differences are due to the fact that asking someone to assess a headline's degree of manipulateness is different from assessing its truth value (e.g., by rating its accuracy or whether it is true or false) because even true information can be presented in a manipulative way (e.g., by using emotionally manipulative language; Brady et al., 2017). However, the same can be argued for the reliability and trustworthiness question framings, both of which behave very similarly to the accuracy and real-vs-fake framings. We encourage further research to gain more insight into this phenomenon, for example by eliciting “top of mind” associations with words such as “manipulativeness”, “reliability” and “accuracy” (see van der Linden, Panagopoulos, & Roozenbeek, 2020).

5.3 Comparing the “integrative” and “classical reasoning” accounts

Finally, we compared two well-known (and somewhat overlapping) accounts of misinformation belief: the “integrative” account (Van Bavel et al., 2021) and the “classical reasoning” account (Pennycook & Rand, 2019; 2020). The former predicts that in addition to analytical skills, actively open-minded thinking (AOT; Baron, 2019) and political ideology (Van Bavel & Pereira, 2018) are consistent predictors of misinformation susceptibility, whereas the latter, in its most extreme form, predicts that analytical thinking (as measured by CRT performance; Pennycook & Rand, 2019) and/or numeracy skills (a second, somewhat different indicator of analytical reasoning ability; Roozenbeek et al., 2020) are more strongly associated with belief in misinformation than political ideology and “myside bias”.

Our results show robust support for the “integrative” account compared to the “classical reasoning” account. Overall, although analytical thinking plays a role, a propensity towards “myside bias” and political conservatism are more strongly correlated with misinformation susceptibility than purely cognitive factors such as numeracy skills and especially

CRT performance. Specifically, actively open-minded thinking is strongly and consistently correlated with misinformation susceptibility. Susceptibility was also consistently higher among those identifying as more politically conservative, indicating that political partisanship plays an important role in misinformation belief (at least in the United States, where our study was conducted). Conversely, performance on a numeracy task and especially analytical thinking ability (as measured by the CRT) were comparatively weakly associated with misinformation susceptibility, although the numeracy task is less clear because many subjects were at ceiling. This means that the correlation between belief in misinformation and analytical thinking may not be robust across different methods of measurement when using a psychometrically validated instrument. These findings are somewhat inconsistent with prior research, which has identified analytical thinking (as measured by CRT performance) as an important predictor of misinformation susceptibility (Pennycook & Rand, 2019; 2020).

Previous research has proposed that both AOT and analytical thinking as measured by the CRT predict reasoning ability, and that both should therefore predict a person's propensity to fall for misinformation, in accordance with the "classical reasoning" account (Erlich et al., 2022). As predictors of misinformation susceptibility, however, AOT and CRT appear to be distinct: we find no collinearity between AOT and CRT (see Figure S5 and Table S23), and, even when removing AOT from the analysis, CRT remains the weakest predictor of veracity discernment ability in all conditions, after political ideology and numeracy skills (see Table S21). Thus, although both the CRT (Frederick, 2005) and AOT (Svedholm-Häkkinen & Lindeman, 2018) scales are measures of reflective thinking ability, they measure distinct constructs within the context of misinformation susceptibility. In fact, these results are consistent with Pennycook, Cheyne, et al. (2020) who find that AOT is negatively correlated with belief in conspiracy theories and specifically that "CRT was a weaker (and often non-significant) predictor for every item relative to either [version of the] AOT-E scale" (p. 487). In line with our findings, the authors therefore correctly conclude that AOT is not merely a proxy for analytical thinking (see also Baron et al., 2015, 2022). One possible explanation for this is that the CRT assesses a participant's ability to correctly solve a set of analytical problems (with correct and incorrect answers), whereas the AOT consists of self-reported agreement about a series of statements about standards for good thinking (such as "people should search actively for reasons why they might be wrong"). In other words, CRT measures cognitive reflection ability (Frederick, 2005; Thomson & Oppenheimer, 2016), whereas AOT is sensitive to "myside bias" (Baron, 2019, p. 10; Svedholm-Häkkinen & Lindeman, 2018, p. 22).

Interestingly, Pennycook, Cheyne et al. (2020) also note that the role of AOT was more pronounced for Democrats than Republicans. For example, they found that higher AOT scores were negatively associated with belief in conspiracy theories for Democrats but that this relationship was not significant for Republicans. In contrast, we find that for both

Democrats and Republicans, AOT is strongly correlated with veracity discernment ability.¹¹

We note several limitations about our study. First, while we made efforts to recruit a large and diverse sample, it was not quota-matched, and we only recruited participants who were United States residents. Importantly, while our sample is well-balanced in terms of gender (with approximately 50% of the sample identifying as female) and US region (Table S1), it is not representative of the US population in terms of age or ethnic/racial background. We do note, however, that the fit values of the binary real-vs-fake condition's SEM were highly similar to those reported by Maertens, Götz, et al. (2022), who made use of several representative samples to assess the validity of the MIST. In addition, Maertens, Götz, et al. (2022) ran studies on different recruitment platforms (Respondi, Prolific, and CloudResearch), as well as in different countries (the United States and the United Kingdom), reporting a high degree of robustness and consistency of the MIST. We thus have good reason to assume that our findings would be highly similar if a representative sample was obtained, or if we had run our study in the UK. Finally, it could be argued that the MIST is not an ecologically valid way of assessing misinformation susceptibility, as the test consists of (partially computer-generated) headlines, without source information, formatting, or other information that would ordinarily accompany a news headline in a real-world environment. We note, however, that the MIST was tested against more ecologically valid item sets such as those used by Pennycook and Rand (2019) and Maertens et al. (2021), showing very strong correlations (Maertens, Götz, et al., 2022). In addition, as the MIST has the advantage (over the currently available more ecologically valid tests) that it is psychometrically validated, we argue that the MIST was the most reliable instrument to use for the present study design.

6 Conclusion

This study is a first attempt at bringing together the large variety of assessment methods used to measure misinformation susceptibility. First of all, we conclude that the use of different question framings and especially response modes should not be expected to yield meaningfully different responses (at least when using the same item set). This finding is of key importance for researchers seeking to compare different studies (e.g., when comparing the efficacy of different anti-misinformation interventions, or for meta-analyses and systematic reviews). We conclude that such comparisons can be safely conducted without a significant risk of similar studies inadvertently assessing fundamentally different constructs. This is good news for the misinformation research community, as there is

¹¹ $r_{\text{Democrats}} = 0.44, p < 0.001$ and $r_{\text{Republicans}} = 0.34, p < 0.001$ (see Table S25). We again note the important role played by political partisanship in predicting misinformation susceptibility: as Table S25 shows, when splitting up the sample by Republicans and Democrats, political ideology is significantly correlated with veracity discernment only for Democrats, but not for Republicans. This implies that more conservative Democrats generally have worse veracity discernment ability than more liberal Democrats, whereas more liberal and more conservative Republicans do not differ in their ability to discern true from false news.

an urgent need to bring structure to the wide variety of approaches, methodologies, and frameworks that have been employed so far. We therefore encourage future work to use our findings as a starting point for further systematising misinformation research.

Second, people's confidence in their primary judgments of true and false news headlines is not meaningfully affected by the question framing or response mode used to elicit this judgment. In addition, primary news headline ratings and confidence ratings measure a similar construct. This opens up the possibility for treating headline ratings as a proxy for confidence: for example, rating a false headline as highly manipulative is a strong indicator of high confidence that it is manipulative. Our findings may therefore act as a bridge to connect the sub-discipline of confidence (and metacognition) to misinformation susceptibility.

Finally, we tested two general approaches to prediction of misinformation susceptibility against each other: the "integrative" account (Van Bavel et al., 2021; van der Linden, 2022), which emphasises the role of myside bias and political partisanship, and the "classical reasoning" account, which argues that a lack of analytical thinking (Pennycook & Rand, 2020, p. 186) is most useful in predicting susceptibility. Our study supports the former over the latter: cognitive factors and analytical thinking (i.e., CRT and numeracy skills) were consistently weaker predictors of belief in misinformation than open-mindedness (i.e., "myside bias") and political ideology. Thus, although cognition and analytical thinking ability can play a role, the ability to consider the viewpoints of those one disagrees with, as well as partisanship and identity-related motivations, appear to be more predictive of misinformation susceptibility. As active open-mindedness was the strongest and most consistent predictor across conditions, we highlight the need to further explore the role of thinking standards as part of an integrated account of misinformation belief.

References

- Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly* 48(2), 409–442. <https://doi.org/10.1086/268840>
- Baron, J. (2019). Actively open-minded thinking in politics. *Cognition*, 188, 8–18. <https://doi.org/10.1016/j.cognition.2018.10.004>
- Baron, J., Isler, O., & Yilmaz, O. (2022). Actively open-minded thinking and the political effects of its absence. *PsyArXiv*. <https://doi.org/10.31234/osf.io/g5jhp>
- Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, 4(3), 265–284. <http://dx.doi.org/10.1016/j.jarmac.2014.09.003>
- Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two

- prebunking interventions against COVID-19 misinformation. *Big Data and Society* 8(1). <https://doi.org/10.1177/20539517211013868>
- Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about Bad News: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, 3(1)(2), 1–9. <https://doi.org/https://doi.org/10.5334/joc.91>
- Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, 121(5), S2-23. <https://doi.org/10.1016/j.amjmed.2008.01.001>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, 6, 149. <https://doi.org/10.3389/fpubh.2018.00149>
- Bradburn, N. (1982). Question-wording effects in surveys. In Hogarth, R.M. (Ed.). *Question framing and response consistency* (pp. 65-76). San Francisco, CA: Jossey-Bass.
- Brady, W.J., Wills, J.A., Jost, J.T., Tucker, J.A., & Van Bavel, J.J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences* 114(28), 7313–7318. <https://doi.org/10.1073/pnas.1618923114>
- Bruine de Bruin, W. (2011). Framing effects in survey design: How respondents make sense of the questions we ask. In G. Keren (Ed.), *Perspectives on Framing* (pp. 303–324). Milton Park: Taylor & Francis.
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412–1427. <https://doi.org/10.1037/pas0000626>
- Compton, J., & Pfau, M. (2005). Inoculation theory of resistance to influence at maturity: recent progress in theory development and application and suggestions for future research. *Annals of the International Communication Association*, 29(1), 97–145. https://doi.org/10.1207/s15567419cy2901_4
- DeCastellarnau, A. (2018). A classification of response scale characteristics that affect data quality: a literature review. *Quality & Quantity* 52(4), 1523–1559. <https://doi.org/10.1007/s11135-017-0533-4>
- Dryhurst, S., Schneider, C. R., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., Spiegelhalter, D., & van der Linden, S. (2020). Risk perceptions of COVID-19 around the world. *Journal of Risk Research*, 23(7–8), 994–1006. <https://doi.org/10.1080/13669877.2020.1758193>
- Erlich, A., Garner, C., Pennycook, G., & Rand, D.G. (2022). Does analytic thinking insulate against pro-Kremlin disinformation? Evidence from Ukraine. *PsyArxiv*. <https://doi.org/10.31234/osf.io/4yrdj>
- Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School (HKS) Misinformation Review* 1(2). <https://doi.org/10.37016/mr-2020-009>
- Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*, 31(4), 439–485. <https://doi.org/10.1006/jrpe.1997.2194>

- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives* 19(4), 25–42. <https://doi.org/10.1257/089533005775196732>
- Gawronski, B. (2021). Partisan bias in the identification of fake news. *Trends in Cognitive Sciences*, 25(9), 723–724. <https://doi.org/10.1016/j.tics.2021.05.001>
- Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, 117(27), 15536 LP – 15545. <https://doi.org/10.1073/pnas.1920498117>
- Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *Journal of General Psychology* 130(2), 149–168. <https://doi.org/10.1080/00221300309601282>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Lee, S. T. H. (2018). Testing for measurement invariance: Does your measure mean the same thing for different participants? *APS Student Notebook*. Found online at: <https://www.psychologicalscience.org/observer/testing-for-measurement-invariance>
- Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour*. <https://doi.org/10.1038/s41562-021-01056-1>
- Maertens, R., Götz, F. M., Schneider, C. R., Roozenbeek, J., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., & van der Linden, S. (2022). *The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment*. PsyArXiv. <https://doi.org/10.31234/osf.io/gk68h>
- Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, 27(1), 1–16. <https://doi.org/10.1037/xap0000315>
- McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press. <https://doi.org/10.4324/9781410601087>.
- McGrew, S. (2020). Learning to evaluate: An intervention in civic online reasoning. *Computers & Education*, 145, 103711. <https://doi.org/https://doi.org/10.1016/j.compedu.2019.103711>.
- Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. *JAMA Internal Medicine*, 173(21), 1952–1958. <https://doi.org/10.1001/jamainternmed.2013.10081>
- Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70, 153–163. <https://doi.org/10.1016/j.jesp.2017.01.006>
- Pehlivanoglu, D., Lin, T., Deceus, F., Heemskerk, A., Ebner, N.C., & Cahill, B.S. (2021). The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive Research: Principles and Implications*, 6(24). <https://doi.org/10.1186/s41235-021-00292-3>
- Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases per-

- ceived accuracy of fake news. *Journal of Experimental Psychology: General*, 147(12), 1865–1880. <https://doi.org/10.1037/xge0000465>
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592, 590–595. <https://doi.org/s41586-021-03344-2>
- Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2020). On the belief that beliefs should change according to evidence: Implications for conspiratorial, moral, paranormal, political, religious, and science beliefs. *Judgment and Decision Making*, 15(4), 476–498. <https://doi.org/10.31234/osf.io/a7k96>
- Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, 31(7), 770–780. <https://doi.org/10.1177/0956797620939054>
- Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality* 88(2), 185–200. <https://doi.org/10.1111/jopy.12476>
- 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. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
- Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. Routledge. <https://doi.org/10.4324/9781315814919>
- Preston, C.C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104(1), 1–15. [https://doi.org/10.1016/s0001-6918\(99\)00050-5](https://doi.org/10.1016/s0001-6918(99)00050-5)
- Price, P. C., & Stone, E. R. (2003). Intuitive evaluation of likelihood judgment producers: Evidence for a confidence heuristic. *Journal of Behavioral Decision Making*, 17(1), 39–57. <https://doi.org/10.1002/bdm.460>
- Revilla, M. A., Saris, W. E., & Krosnick, J. A. (2014). Choosing the number of categories in agree–disagree scales. *Sociological Methods & Research* 43(1), 73–97. <https://doi.org/10.1177/0049124113509605>
- Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy nudges? A pre-registered direct replication of Pennycook et al. (2020). *Psychological Science* 32(7), 1–10. <https://doi.org/10.1177/09567976211024535>
- Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). Disentangling item and testing effects in inoculation research on online misinformation. *Educational and Psychological Measurement*, 81(2), 340–362. <https://doi.org/10.1177/0013164420940378>
- Roozenbeek, J., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, 7(201199). <https://doi.org/10.1098/rsos.201199>

- Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Humanities and Social Sciences Communications*, 5(65), 1–10. <https://doi.org/10.1057/s41599-019-0279-9>
- Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (under review). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*.
- Saleh, N., Roozenbeek, J., Makki, F., McClanahan, W., & van der Linden, S. (2021). Active inoculation boosts attitudinal resistance against extremist persuasion techniques – A novel approach towards the prevention of violent extremism. *Behavioural Public Policy*, 1–24. <https://doi.org/10.1017/bpp.2020.60>
- Salovich, N. A., & Rapp, D. N. (2020). Misinformed and unaware? Metacognition and the influence of inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. <https://doi.org/10.1037/xlm0000977>
- Schumacker, R. E., Lomax, R. G., & Schumacker, R. (2015). *A beginner's guide to structural equation modeling* (4th ed.). Routledge. <https://www.routledge.com/A-Beginners-Guide-to-Structural-Equation-Modeling-Fourth-Edition/Schumacker-Lomax-Schumacker-Lomax/p/book/9781138811935>
- Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54(2), 93–105. <https://doi.org/10.1037/0003-066X.54.2.93>
- Schwartz, L. M. L., Woloshin, S. S., Black, W. C. W., & Welch, H. G. H. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, 127, 966–972. <https://doi.org/10.7326/0003-4819-127-11-199712010-00003>
- Smith, T. W. (1995). A Review: The Holocaust Denial Controversy. *Public Opinion Quarterly* 59(2), 269–295. <https://www.jstor.org/stable/2749705>
- Stanovich, K. E., West, R. F., & Toplak, M. E. (2013). Myside Bias, Rational Thinking, and Intelligence. *Current Directions in Psychological Science*, 22(4), 259–264. <https://doi.org/10.1177/0963721413480174>
- Svedholm-Häkkinen, A. M., & Lindeman, M. (2018). Actively open-minded thinking: development of a shortened scale and disentangling attitudes towards knowledge and people. *Thinking & Reasoning*, 24(1), 21–40. <https://doi.org/10.1080/13546783.2017.1378723>
- Swire, B., Berinsky, A., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society Open Science* 4(3): 160802. <https://doi.org/10.1098/rsos.160802>
- Tenney, E. R., Spellman, B. A., & MacCoun, R. J. (2008). The benefits of knowing what you know (and what you don't): How calibration affects credibility. *Journal of Experimental Social Psychology*, 44(5), 1368–1375. <https://doi.org/j.jesp.2008.04.006>
- Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, 11(1), 99–113. <http://journal.sjdm.org/15/151029/jdm151029.pdf>
- Tormala, Z. L., & Petty, R. E. (2004). Source credibility and attitude certainty: A metacognitive analysis of resistance to persuasion. *Journal of Consumer Psychology*, 14(4), 427–442. https://doi.org/https://doi.org/10.1207/s15327663jcp1404_11

- Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by repetition: Explanations and implications. *Current Directions in Psychological Science*, 28(3), 247–253. <https://doi.org/10.1177/0963721419827854>
- Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Haslam, S. A., Jetten, J., . . . Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, 4(5), 460–471. <https://doi.org/10.1038/s41562-020-0884-z>
- Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., Tucker, J. A. (2021). Political psychology in the digital (mis)information age: A model of news belief and sharing. *Social Issues and Policy Review* 15(1), 84–113. <https://doi.org/10.1111/sipr.12077>
- Van Bavel, J. J., & Pereira, A. (2018). The partisan brain: An identity-based model of political belief. *Trends in Cognitive Sciences*, 22(3), 213–224. <https://doi.org/j.tics.2018.01.004>
- Van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*. <https://doi.org/10.1038/s41591-022-01713-6>
- Van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: the emergence of political bias in perceptions of fake news. *Media, Culture & Society* 42(3), 460–470. <https://doi.org/10.1177/0163443720906992>
- Van der Linden, S., & Roozenbeek, J. (2020). Psychological inoculation against fake news. In R. Greifeneder, M. Jaffé, E. Newman, & N. Schwartz (Eds.), *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation* (pp. 147-170). London: Psychology Press.
- Van der Linden, S., Roozenbeek, J., Maertens, R., Basol, M., Kácha, O., Rathje, S., & Steenbuch Traberg, C. (2021). How can psychological science help counter the spread of fake news? *Spanish Journal of Psychology*, 24, 1–9. <https://doi.org/10.1017/SJP.2021.23>
- Xu, H., & Tracey, T.J.G. (2017). Use of multi-group confirmatory factor analysis in examining measurement invariance in counseling psychology research. *European Journal of Counseling Psychology* 6(1), 75–82. <https://doi.org/964/ejcop.v6i1.120>
- Zillmann, D., Gibson, R., & Sargent, S.L. (1999). Effects of photographs in news-magazine reports on issue perception. *Media Psychology*, 1(3), 207–228. https://doi.org/10.1207/s1532785xmep0103_2

APPENDIX: Items used in MIST20 (also shown in Table S2 of the [Supplement](#)). * indicates inclusion in MIST8. p indicates possibly right-leaning items excluded in one analysis. F indicates Fake; R indicates Real.

-
- F1* Government Officials Have Manipulated Stock Prices to Hide Scandals
 - F2 The Corporate Media Is Controlled by the Military-industrial Complex: The Major Oil Companies Own the Media and Control Their Agenda
 - F3* New Study: Left-Wingers Are More Likely to Lie to Get a Higher Salary
 - F4p The Government Is Manipulating the Public's Perception of Genetic Engineering in Order to Make People More Accepting of Such Techniques
 - F5p Left-Wing Extremism Causes 'More Damage' to World Than Terrorism, Says UN Report
 - F6* Certain Vaccines Are Loaded with Dangerous Chemicals and Toxins
 - F7 New Study: Clear Relationship Between Eye Color and Intelligence
 - F8* The Government Is Knowingly Spreading Disease Through the Airwaves and Food Supply
 - F9 Ebola Virus 'Caused by US Nuclear Weapons Testing', New Study Says
 - F10p Government Officials Have Illegally Manipulated the Weather to Cause Devastating Storms
 - R1* Attitudes Toward EU Are Largely Positive, Both Within Europe and Outside It
 - R2 One-in-Three Worldwide Lack Confidence in NGOs
 - R3 Reflecting a Demographic Shift, 109 US Counties Have Become Majority Nonwhite Since 2000
 - R4 International Relations Experts and US Public Agree: America Is Less Respected Globally
 - R5* Hyatt Will Remove Small Bottles from Hotel Bathrooms by 2021
 - R6 Morocco's King Appoints Committee Chief to Fight Poverty and Inequality
 - R7* Republicans Divided in Views of Trump's Conduct, Democrats Are Broadly Critical
 - R8 Democrats More Supportive than Republicans of Federal Spending for Scientific Research
 - R9* Global Warming Age Gap: Younger Americans Most Worried
 - R10 US Support for Legal Marijuana Steady in Past Year
-