Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Psychology

Review

Mean rating difference scores are poor measures of discernment: The role of response criteria

Philip A. Higham, Ariana Modirrousta-Galian and Tina Seabrooke

Abstract

Many interventions aim to protect people from misinformation. Here, we review common measures used to assess their efficacy. Some measures only assess the target behavior (e.g., ability to spot misinformation) and therefore cannot determine whether interventions have overly general effects (e.g., erroneously identifying accurate information as misinformation). Better measures assess discernment, the ability to discriminate target from non-target content. Discernment can determine whether interventions are overly general but is often measured by comparing differences in mean ratings between target and non-target content. We show how this measure is confounded by the configuration of response criteria, leading researchers to incorrectly conclude that an intervention improves discernment. We recommend using measures from signal detection theory, such as the area under the receiver operating characteristic curve, to assess discernment.

Addresses

University of Southampton, Southampton, United Kingdom

Corresponding author: Higham, Philip A (higham@soton.ac.uk)

Keywords

Misinformation, Receiver operating characteristic analysis, Signal detection theory, Area under the curve, Mean rating difference scores.

Introduction

Misinformation on the internet has become a major problem in modern society. To combat its influence, a number of interventions have been introduced, such as those aimed at reducing belief in misinformation, reducing misinformation sharing, and improving people's ability to spot manipulative techniques in social media posts (e.g., Refs. [1−8]; see Refs. [9,10] for reviews). To assess whether these interventions work, different measures have been proposed which vary in their quality (e.g., Refs. [11−18]). In this article, we review some of the more common measures, assess their suitability, and suggest an alternative measure that overcomes the limitations of those more commonly used.

Successful behavioral interventions in any domain ideally should be specific, affecting the targeted behavior and having little influence on any other behaviors [19,20]. For example, when evaluating the effectiveness of interventions that moderate the content of toxic online communities, it is vital to not just assess changes within the moderated communities, but also the changes in other communities to which users might have migrated [21]. Thus, whatever measure is used to assess an intervention should evaluate not just the effects on the targeted behavior, but non-target behaviors as well. However, these guidelines have not always been followed in misinformation research (e.g., see Refs. [22−25]). For example, some recent research has examined how short videos on YouTube can be used to inoculate people against manipulative techniques used to create misleading content (e.g., [5], Study 7, [26]). In [[5], Study 7], participants viewed an inoculation video that explained a manipulative technique and provided an example (e.g., false dichotomies). Later, participants were presented with a statement that used the technique and asked to identify it from a list of alternatives. Participants who watched the video were better able to identify the technique compared to participants who had not. However, this design cannot speak to the specificity of the video intervention because statements that did not include manipulative content were not included. If participants also identified manipulative content in such statements, the effect of the intervention would be too broad, leading people to see manipulative content when it was not there. Conversely, if the rate of identifying manipulative content in statements that had none was no different (or less) in the video condition versus the no-video condition, then the intervention could be deemed successful.

Other research has included the necessary controls to assess the generality of misinformation interventions

Current Opinion in Psychology 2024, **56**:101785

This review comes from a themed issue on **The Psychology of Misinformation 2024**

Edited by **Gordon Pennycook** and **Lisa K. Fazio**

For complete overview about the section, refer Generation COVID: Coming of Age Amid the Pandemic (2024)

Available online 19 December 2023

https://doi.org/10.1016/j.copsyc.2023.101785

(e.g., Refs. [27–33]). For example, Roozenbeek et al. [34] (see also [6]) had participants rate the reliability (on a 7-point scale ranging from 1 = *unreliable* to 7 = *reliable*) of nine social media posts before and after playing Bad News, an internet game intended to inoculate players against fake news. Seven of the posts were fake whereas the remaining two were true. They found that mean ratings to both true and fake news decreased from pre-test to post-test, but that the difference was greater for fake news, suggesting that the game had (fairly) specific effects. This process of comparing the difference in mean ratings (per participant) between a pretest and a post-test (or between a control condition and experimental condition) for true and fake news items is referred to as *veracity discernment*. It is a common methodology in the literature, whether that methodology uses multiple *t*-tests, ANOVA, or OLS [18]. However, comparing mean difference scores to measure discernment is problematic, which may explain why some studies have concluded that gamified interventions improve veracity discernment [27,30,34–36], whereas others have not [16,28,31,37] (see Refs. [17,38] for meta-analyses). For the remainder of this review, we (1) elaborate on the problems that comparing mean ratings introduce; (2) offer an alternative method of evaluating discernment; and (3) identify a case in the literature where analysis of difference scores has led to misleading conclusions, a problem that does not occur if the alternative method is used instead.

Measuring discernment with mean difference scores: The problem

Comparing mean ratings between experimental conditions may seem like a neutral procedure. However, when examined more closely, it becomes apparent that the procedure conflates different processes that underly decision-making. To understand this conflation, it is fruitful to consider formalized theories of decision-making such as signal detection theory (SDT). While SDT is routinely used to analyze data in some areas of psychology (e.g., recognition memory; [39]), its application to misinformation research has been limited (although see Refs. [14,16,17,40–43]). SDT assumes that there are two separate, measurable processes involved in any discernment task. The first process relates to whether an intervention affects how people subjectively assess different classes of stimuli such as true versus fake headlines. For example, after an inoculation intervention, do true news and fake news items subjectively seem more true and/or more fake, respectively? The second process is a nuisance variable, which has nothing to do with whether an intervention affects subjective assessment but rather with how the rating scale is used. In some contexts, people might assign a rating that is higher or lower than in another context even though their subjective assessment has not changed. This is known as *criteria setting*. Critically, analysis of mean ratings makes no distinction between these separate processes. Instead, mean ratings are usually interpreted to directly reflect subjective assessment with no consideration given to the influence of criteria setting.

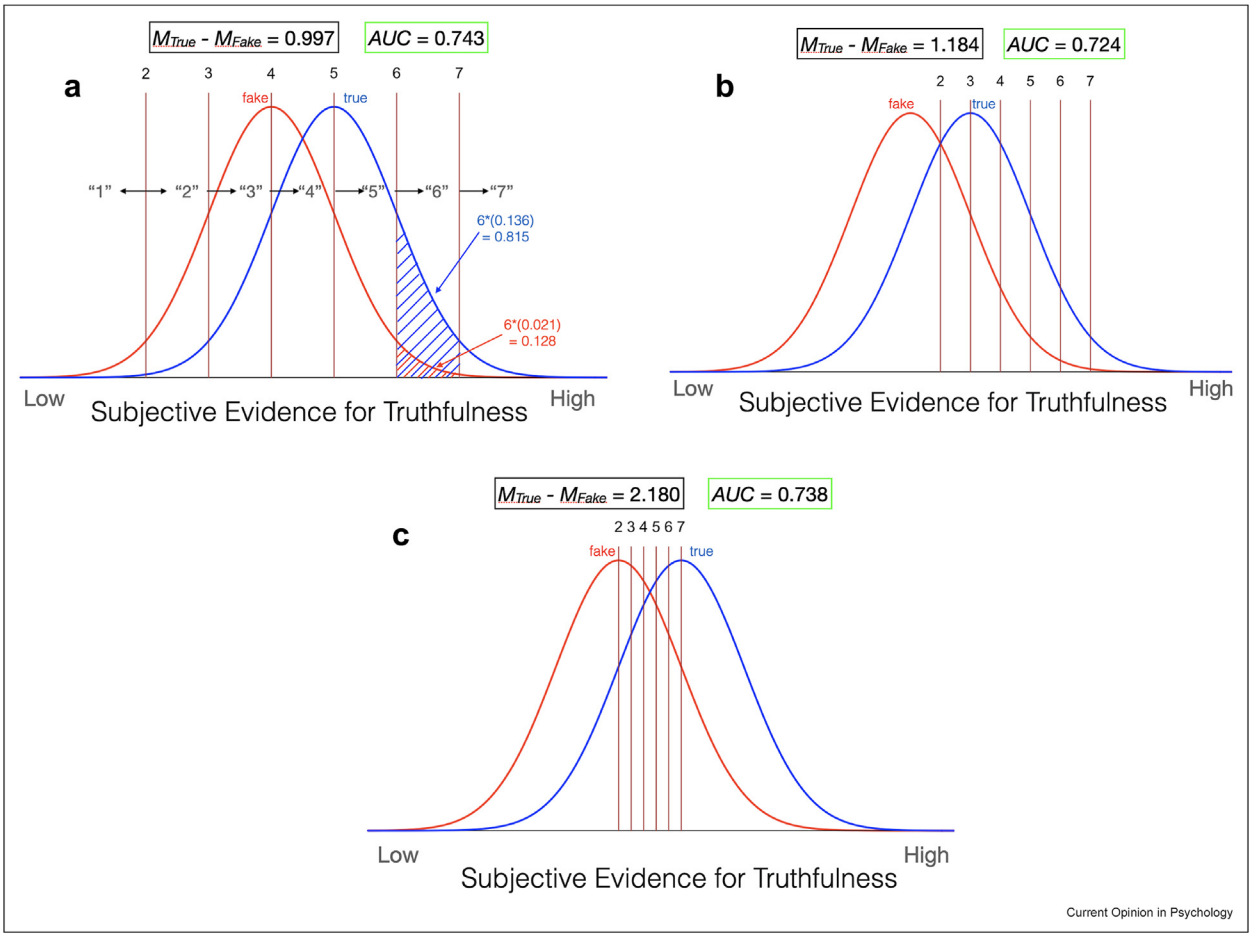
A real-world example might help to clarify the distinction. The lead author spent the early part of his academic career in Canada where the average mark assigned to undergraduate student essays was about 70%. After moving to the UK, he used the same standard when marking his first batch of essays. However, it quickly became apparent that the grades he had assigned were significantly inflated compared to the marks of other British academics because 70% is a first-class mark in the UK. Consequently, he reassessed the essays and his marks, aiming to achieve a reduced mean mark of about 62%, which is more typical of the UK standard.

An important point here is that the change to the mean marks following reassessment *was not due to a change in the subjective evaluation of the essays*; a poor essay was still considered poor, and a good essay was still considered good. What changed was the way he *assigned scale values to his internal subjective assessments* of essay quality. In other words, he set more stringent criteria to assign marks at different levels (e.g., only excellent essays achieved 70%). Thus, subjective evaluation can be unaffected by some intervention (such as remarking), but different criterion settings can affect mean scores. Critically, there is no way to know whether changes to subjective evaluation, criterion setting, or both have produced the observed differences to mean ratings. In other words, the effects of subjective evaluation and criterion setting on mean ratings are *confounded*.

This confounding of subjective evaluation and criterion setting with mean ratings clearly poses problems for those studies that have only examined the effect of an intervention on one type of stimulus, such as fake news. If the intervention lowered reliability ratings, it is entirely ambiguous as to whether that reduction was due to improved subjective evaluation of fake news or to more stringent response criteria. However, what about studies that have investigated *discernment* as measured by *differences* between mean ratings of true and fake news items? If the change to mean ratings following an intervention was due only to shifting criteria, perhaps mean ratings of the two item types would decrease (or increase) to the same extent. Thus, when a difference score is taken, it would remain unaltered, accurately reflecting no change to discernment.

Unfortunately, it is straightforward to show that this is not the case: *both* mean ratings *and* discernment (as measured by difference scores) are profoundly affected by criterion setting. To understand why, it is useful to

Figure 1



Signal Detection Models of True and Fake News Discernment With a 7-Point Scale. Participants Adopt Six Criteria on an Underlying Subjective Dimension of Truthfulness and Use Them to Assign 7 Scale Values (1–7) to True and Fake New Items, Normally Distributed Over a Subjective Dimension of Truthfulness. Because Participants can Control the Criteria They Set, Many Different Configurations are Possible. The Criteria in Panels A and C Represent Different Unbiased Criteria Configurations, but With Different Dispersions. Panel B Shows a Conservative Configuration.
Note. To compute mean ratings, each scale value is weighted by the relevant areas under the true and fake item distributions and summed across the whole scale. One part of the process is shown in Panel A for scale value "6". The full computation is shown in Table 1. An ideal measure of discernment would not vary as the criteria configurations change across the three panels because discernment is constant (i.e., the overlap of the distributions is constant). However, the difference in mean ratings changes considerably, showing that it is a poor measure of discernment because it is contaminated by criteria placement. For Panel A, M_{True} = 4.499; M_{Fake} = 3.501; $\mathbf{M_{Diff}}$ = **0.997**. For Panel B, M_{True} = 2.748; M_{Fake} = 1.564; $\mathbf{M_{Diff}}$ = **1.184**. For Panel C, M_{True} = 5.090; M_{Fake} = 2.910; $\mathbf{M_{Diff}}$ = **2.180**. Of particular concern is that Panels A and C have criteria configurations that are unbiased; that is, they are centered around the intersection point of the two curves. They differ only in dispersion. In Panel A, participants are using all the scale values approximately equally, whereas in Panel C, participants are assigning mostly 1s and 7s. This trivial difference, which has nothing to do with discernment, creates more than a twofold difference in mean difference scores.

use a signal detection model that can separate subjective evaluation and criterion setting such as the models shown in Figure 1. Instead of essay quality, there is a horizontal dimension of *subjective evidence of truthfulness (SET)* ranging from little evidence on the left to a lot of evidence on the right. As with essay grading, people are assumed to set criteria for assigning scale values.[1] These are shown as vertical lines placed on the underlying SET dimension, with higher scale values further up the dimension than lower scale values. Also shown in Figure 1 are two normal distributions that represent the probability of different levels of subjective truthfulness for the true (blue) and fake items (red) that participants are rating. The distributions have different means such that fake items have lower SET than true items. Importantly, *the separation of the distributions represents discernment*. If the distributions were completely overlapping versus completely separated, discernment would be at chance versus perfect, respectively. Across

[1] In this case, participants are using a 1–7 scale instead of a 0–100 scale typical of essay marking, although any scale with any number of levels can be modeled in the same way.

the three panels of Figure 1, discernment is moderate and constant, with the true and fake news distributions separated by one *SD* unit (z-score).

To assign scale values, items with SET below the "2" criterion are assigned "1". Those with SET between the "2" and "3" criteria are assigned "2", and so on up to "7" (see Panel A of Figure 1). Whereas the three panels in Figure 1 have constant discernment, they have different criteria configurations. Panel A represents unbiased criteria which are widely dispersed, indicating that participants are using all the points on the scale (1–7). They are unbiased because the midpoint of the criteria configuration (halfway between scale values 4 and 5) is placed at the intersection point of the two distributions.[2] Panel B represents a conservative configuration, with the criteria set to the right of the intersection point. Panel C represents another unbiased configuration but with little dispersion, indicating that participants are using the extremes of the scale (1 and 7) extensively.

Note that this way of representing the decision-making process separates discernment (the amount of overlap of the fake and true item distributions) and scale usage (the positioning of the response criteria). Thus, it is now possible to determine how changes in response configurations between the different panels affect mean differences. To compute mean true- and fake-news ratings, each scale value is weighted by the corresponding area and summed over the whole scale. This computation is depicted graphically for scale value "6" in Panel A of Figure 1. The full computation is shown in Table 1.

The difference between the means for true and fake items is shown in a box with a black border in each panel of Figure 1. Because only the configurations of the response criteria change between the panels, whereas the overlap of the distributions (which reflects discernment) is static, a good discernment measure should not vary between the panels. However, it is clear to see that this is not the case for the mean difference measure. Moving the criteria from the unbiased configuration in Panel A to a more conservative position in Panel B increases the difference score from 0.997 to 1.184. Worse, keeping the criteria configuration unbiased but reducing the dispersion in Panel C compared to Panel A more than doubles the difference score to 2.180.

Table 1

Complete computation of mean ratings for true and fake news items for the signal detection model shown in Panel A of Figure 1.

True News Distribution								
Scale (S)	1	2	3	4	5	6	7	Sum
Area (A)	0.001	0.021	0.136	0.341	0.341	0.136	0.023	1.000
S X A	0.001	0.043	0.408	1.365	1.707	0.815	0.159	4.499
Fake News Distribution								
Scale (S)	1	2	3	4	5	6	7	Sum
Area (A)	0.023	0.136	0.341	0.341	0.136	0.021	0.001	1.000
S X A	0.023	0.272	1.024	1.365	0.680	0.128	0.009	3.501

Note. Mean ratings are computed by multiplying the relevant area of each distribution by the corresponding scale value and summing the products over the whole distribution. If difference scores are used, then discernment would be the mean for true news (4.499) minus the mean for fake news (3.501), which is equal to 0.997.

Receiver operating characteristic (ROC) analysis: The solution

Clearly, the situation depicted in Figure 1 is unacceptable. Irrelevant factors such as the tendency to use the full scale or only the extreme values, or to require more versus less evidence before assigning particular scale values, should not have such a profound effect on the discernment estimate. However, such outcomes are inevitable if mean differences are used as the measure of discernment. Fortunately, the remedy is straightforward: use a measure derived from the receiver operating characteristic (ROC) curve, which is based on SDT, rather than mean differences (e.g., Ref. [44]).

To create an ROC curve, the scale values themselves (e.g., 1–7) are ignored. The scale values serve to define the placements of the different criteria, as in Figure 1, but unlike mean ratings, they are not included in the calculation of discernment. Thus, a 1–7 scale would yield the same ROC curve as a 0–6 scale if the criteria corresponding to each of the scale values are in the same place on the SET dimension. Instead, the ROC curve uses the criteria to produce a series of (x, y) points reflecting the cumulative areas under the true (y) and fake (x) news item distributions. By convention, the x-axis points and y-axis points are referred to as *false alarm rates* (FARs) and *hit rates* (HRs), respectively.

To create the curve, first, the areas above the highest scale value for each distribution are computed.[3] Next, the areas under the fake and true news distributions between the second highest and the highest criterion are added to the cumulative total to create a new (FAR, HR) point. Then, the area between the third highest and second highest criterion is added to create a third (FAR, HR) point, and so on until all the criteria have produced points. Finally, the remaining areas below the lowest criterion are added to the cumulative total to yield the point (1, 1). These points are plotted on an

[2] A criterion at the intersection point maximizes response accuracy.

[3] Empirically, these areas (and the other areas involved in the calculation) are based on the proportion of true and fake news items that are assigned particular scale values. For example, if 5 of 20 true news items and 2 of 20 fake news items were each assigned "7", then the areas above the "7" criterion would be 0.25 and 0.05, respectively.

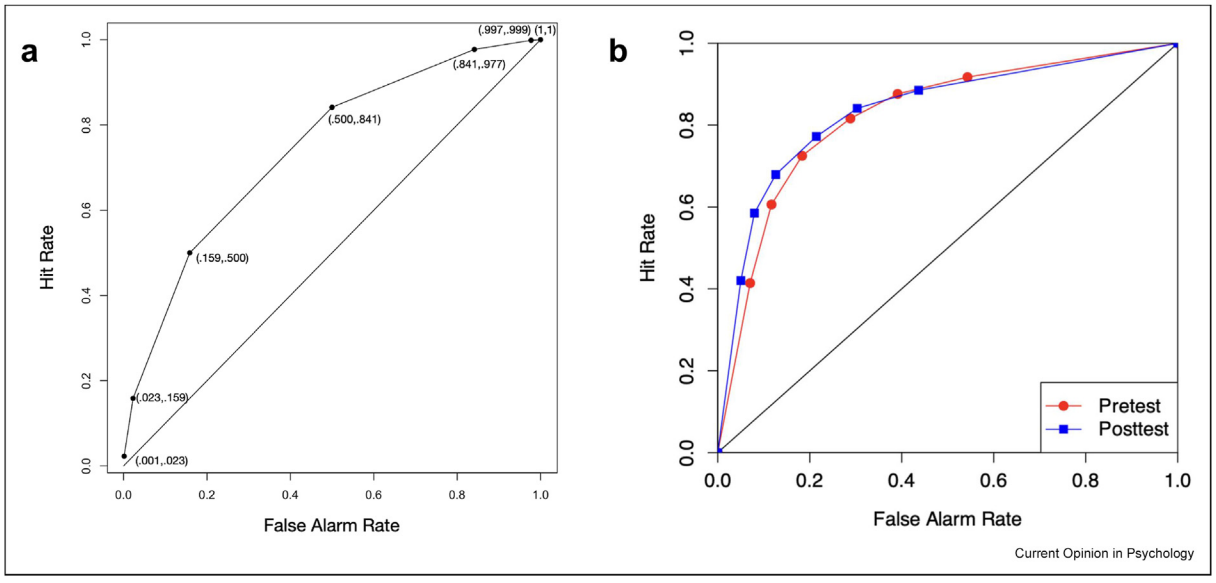
Table 2

Areas and cumulative areas between scale values for the true and fake news item distributions for the model represented in Panel A of Figure 1.

True News Distribution							
Scale Value	1	2	3	4	5	6	7
Area	0.001	0.021	0.136	0.341	0.341	0.136	0.023
Cumulative Area (HRs)	1.000	0.999	0.977	0.841	0.500	0.159	0.023
Fake News Distribution							
Scale Value	1	2	3	4	5	6	7
Area	0.023	0.136	0.341	0.341	0.136	0.021	0.001
Cumulative Area (FARs)	1.000	0.977	0.841	0.500	0.159	0.023	0.001

Note. The cumulative areas correspond to the model in Panel A of Figure 1 and they are plotted on the receiver operating characteristic (ROC) curve in Panel A of Figure 2. HR = hit rate; FAR = false alarm rate.

Figure 2



Receiver Operating Characteristic (ROC) Curves. The Curve in Panel A Corresponds to the Cumulative Areas in Table 2 and the Model in Panel A of Figure 1. The Curves in Panel B corresponds to the Pre-Test and Post-Test in Roozenbeek et al.'s (2022) Experiment 1.

ROC curve with FARs on the x-axis and HRs on the y-axis. The computations corresponding to the model shown in Panel A of Figure 1 are indicated in Table 2, and the ROC curve corresponding to the same model is displayed in Panel A of Figure 2.

Note that the ROC curve in Panel A of Figure 2 bows upward from the chance diagonal, that is, the straight line extending from (0, 0) to (1, 1).[4] This bowing is indicative of discernment and can be indexed by computing the area under the ROC curve (AUC). The simplest way to compute AUC is to create a series of trapezoids by drawing straight lines between points and summing them (e.g., Refs. [45–48]). Critically, note that the extent to which the curve bows from the diagonal is not necessarily impacted by different criteria configurations. Different configurations simply vary the location of the points on the curve, not the extent of bowing. Conservative criteria, for example, will cause points to cluster in the bottom-left of the curve, whereas liberal criteria will cause points to cluster in the top-right. Similarly, extensive use of extreme scale values will result in points that are clustered together, whereas equal use of all scale values will spread the points apart.

[4] The chance diagonal corresponds to a model where the HRs and FARs are equal for all criteria, which indicates that discernment is at chance (i.e., complete overlap of the distributions).

Thus, unlike mean differences, AUC is unlikely to be affected by different criteria configurations to the same extent. Indeed, this conclusion was confirmed for the three models shown in Figure 1; the AUC values are shown in the boxes with green borders. Note that AUC varied over a range of 0.019, while mean differences scores varied over a range of 1.183, which is more than 62 times greater. This result shows that, unlike mean difference scores, AUC is mostly unaffected by different criteria configurations, making it a superior measure of discernment.

Misleading conclusions due to criteria shifts: An example from the literature

Section 1.2.2 showed that changes to criteria configurations alone can have a large effect on discernment when it is measured with difference scores. This finding is concerning because, rather than the criteria being fixed, [17], showed that inoculation interventions can cause response criteria to shift. Potentially, this criteria shift could lead researchers to conclude that an intervention has improved discernment when, in fact, it has simply caused changes to the criteria configuration. To explore this possibility, we re-examined the seven papers reanalyzed in Ref. [17], and found that Experiment 1 in Ref. [34] conformed to this scenario.[5] Compared to mean pre-test veracity ratings of true and fake headlines, the authors reported that the Bad News game decreased mean ratings to both true and fake news on a post-test, but that the change was larger for fake news, thereby increasing discernment.[6] We reanalyzed the raw data with ROC analysis (see Panel B of Figure 2 for the corresponding ROC curve) and found that AUC did not significantly differ between the pre-test ($M = 0.82$, $SD = 0.22$) and the post-test ($M = 0.83$, $SD = 0.22$), $t(1{,}215) = 0.76$, $p = .448$, $d = 0.02$, 95% CI [−0.01, 0.02], $BF_{10} = 0.04$. However, mean $B''D$ significantly increased between the pre-test ($M = 0.11$, $SD = 0.64$) and the post-test ($M = 0.25$, $SD = 0.62$), $t(1{,}215) = 11.69$, $p < .001$, $d = 0.37$, 95% CI [0.12, 0.16], $BF_{10} = 3.53 \times 10^{26}$.[7] In other words, when discernment was measured with a bias-free index (AUC), Bad News had no effect on discernment. It was only if discernment was measured with difference scores that are confounded with criteria configurations that Bad News had an effect.

[5] Out of the seven papers reanalyzed in Ref. [17], two ([29,34]) calculated mean difference scores as a measure of discernment, and both showed instances where difference scores indicated improved discernment whereas ROC analysis indicated a conservative shift of the response criteria.

[6] We discovered when analyzing the data that the means included in the supplementary materials for this study were computed incorrectly. However, we were assured by the authors that once the errors were corrected, the main conclusions were unaltered (i.e., bigger mean decrease to fake news ratings than true news ratings in the post-test compared to the pre-test). An erratum has just been published at the time of writing [49]. Our analyses were based on the corrected data.

[7] B″D is a non-parametric measure of criterion placement. Like mean C, negative versus positive values indicate liberal versus conservative criterion placement, respectively. For more detail, see Ref. [50].

Conclusions and recommendations

Overall, the results of our analyses suggest that researchers should avoid intuitive measures such as differences in mean ratings when measuring discernment. Instead, they should use bias-free measures of discernment such as AUC which is based on ROC analysis. Only by using a tool such as ROC analysis can researchers avoid mistaking differences in criteria configurations for differences in discernment. For a worked example of how to conduct ROC analysis in R, which includes plotting ROC curves as well as calculating AUC (discernment) and $B''D$ (response bias), see our supplementary materials (https://osf.io/x8z9d/). The supplementary materials also include all the analytic codes we used for the reanalysis in section 1.3.

Author contributions

Philip A. Higham: Conceptualization, Formal analysis, Writing − original draft. **Ariana Modirrousta-Galian**: Conceptualization, Formal analysis, Writing − review & editing**. Tina Seabrooke**: Conceptualization, Writing − review & editing.

Funding

This work was supported in part by an Economic and Social Research Council South Coast Doctoral Training Partnership studentship (ES/P000673/1) awarded to Ariana Modirrousta-Galian.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and R code used in the manuscript can be found on https://osf.io/x8z9d/.

References

References of particular interest have been highlighted as:

* of special interest
** of outstanding interest

1. Basol M, Roozenbeek J, Van der Linden S: **Good news about bad news: gamified inoculation boosts confidence and cognitive immunity against fake news**. *J. Cogn.* Jan. 2020, **3**:2, https://doi.org/10.5334/joc.91.

2. Cook J, Lewandowsky S, Ecker UKH: **Neutralizing misinformation through inoculation: exposing misleading argumentation techniques reduces their influence**. *PLoS One* May 2017, **12**, e0175799, https://doi.org/10.1371/journal.pone.0175799.

3. Pennycook G, Rand DG: **Fighting misinformation on social media using crowdsourced judgments of news source quality**. *Proc Natl Acad Sci USA* Feb. 2019, **116**:2521−2526, https://doi.org/10.1073/pnas.1806781116.

4. Pennycook G, Rand DG: **Accuracy prompts are a replicable and generalizable approach for reducing the spread of**

misinformation. *Nat Commun* Apr. 2022, **13**:2333, https://doi.org/10.1038/s41467-022-30073-5.

5. Roozenbeek J, Van Der Linden S, Goldberg B, Rathje S, Lewandowsky S: **Psychological inoculation improves resilience against misinformation on social media**. *Sci Adv* Aug. 2022, **8**:eabo6254, https://doi.org/10.1126/sciadv.abo6254.

6. Roozenbeek J, van der Linden S: **Fake news game confers psychological resistance against online misinformation**. *Palgrave Commun* 2019, **5**, https://doi.org/10.1057/s41599-019-0279-9.

7. Moore RC, Hancock JT: **A digital media literacy intervention for older adults improves resilience to fake news**. *Sci Rep* Apr. 2022, **12**:6008, https://doi.org/10.1038/s41598-022-08437-0.

8. Hameleers M: **Separating truth from lies: comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands**. *Inf Commun Soc* Jan. 2022, **25**:110–126, https://doi.org/10.1080/1369118X.2020.1764603.

9. Ecker UKH, Lewandowsky S, Cook J, Schmid P, Fazio LK, Brashier N, Kendeou P, Vraga EK, Amazeen MA: **The psychological drivers of misinformation belief and its resistance to correction**. *Nat. Rev. Psychol.* Jan. 2022, **1**:13–29, https://doi.org/10.1038/s44159-021-00006-y.

10. Lewandowsky S, van der Linden S: **Countering misinformation and fake news through inoculation and prebunking**. *Eur Rev Soc Psychol* Jul. 2021, **32**:348–384, https://doi.org/10.1080/10463283.2021.1876983.

11. Lees J, Banas JA, Linvill D, Meirick PC, Warren P: **The *Spot the Troll Quiz* game increases accuracy in discerning between real and inauthentic social media accounts**. *PNAS Nexus* Apr. 2023, **2**:pgad094, https://doi.org/10.1093/pnasnexus/pgad094.

12. Orosz G, Paskuj B, Faragó L, Krekó P: **A prosocial fake news intervention with durable effects**. *Sci Rep* Mar. 2023, **13**:3958, https://doi.org/10.1038/s41598-023-30867-7.

13. Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG: **Shifting attention to accuracy can reduce misinformation online**. *Nature* Apr. 2021, **592**:590–595, https://doi.org/10.1038/s41586-021-03344-2.

14. ** Batailler C, Brannon SM, Teas PE, Gawronski B: **A signal detection approach to understanding the identification of fake news**. *Perspect Psychol Sci* Jul. 2021, 174569162098613, https://doi.org/10.1177/1745691620986135.
By using signal detection theory (SDT) to reanalyze two existing data sets, the authors provide further insight into how partisan bias, cognitive reflection, and prior exposure affect fake news detection. This paper was the first to focus specifically on SDT as a valuable tool for analyzing data from misinformation research.

15. Pennycook G, McPhetres J, Zhang Y, Lu JG, Rand DG: **Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy-nudge intervention**. *Psychol Sci* Jul. 2020, **31**:770–780, https://doi.org/10.1177/0956797620939054.

16. Modirrousta-Galian A, Higham PA, Seabrooke T: **Effects of inductive learning and gamification on news veracity discernment**. *J. Exp. Psychol. Appl.* 2023, https://doi.org/10.1037/xap0000458.

17. Modirrousta-Galian A, Higham PA: **Gamified inoculation interventions do not improve discrimination between true and fake news: reanalyzing existing research with receiver operating characteristic analysis**. *J. Exp. Psychol. Gen.* 2023, https://doi.org/10.1037/xge0001395.

18. * Guay B, Berinsky AJ, Pennycook G, Rand D: **How to think about whether misinformation interventions work**. *Nat. Hum. Behav.* 2023, https://doi.org/10.1038/s41562-023-01667-w.
The authors discuss how researchers should determine the effectiveness of misinformation interventions. They recommend measuring the effects of misinformation interventions on people's ability to discern between true and false content.

19. * Van Der Meer TGLA, Hameleers M, Ohme J: **Can fighting misinformation have a negative spillover effect? How warnings for the threat of misinformation can decrease general news credibility**. *Journal Stud* Apr. 2023, **24**:803–823, https://doi.org/10.1080/1461670X.2023.2187652.
The authors studied the potential negative spill over effect of attempts to fight misinformation. They found that misinformation warnings decreased participants' credibility ratings of authentic news.

20. Clayton K, *et al.*: **Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media**. *Polit Behav* Dec. 2020, **42**:1073–1095, https://doi.org/10.1007/s11109-019-09533-0.

21. Trujillo A, Cresci S: **Make Reddit great again: assessing community effects of moderation interventions on r/The_Donald**. *Proc. ACM Hum.-Comput. Interact.* Nov. 2022, **6**:1–28, https://doi.org/10.1145/3555639. CSCW2.

22. Porter E, Wood TJ: **The global effectiveness of fact-checking: evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom**. *Proc Natl Acad Sci USA* Sep. 2021, **118**, e2104235118, https://doi.org/10.1073/pnas.2104235118.

23. Andő S, Akesson J: **Nudging away false news: evidence from a social norms experiment**. *Digit. Journal.* Jan. 2021, **9**:106–125, https://doi.org/10.1080/21670811.2020.1847674.

24. Roozenbeek J, Van Der Linden S: **The fake news game: actively inoculating against the risk of misinformation**. *J Risk Res* May 2019, **22**:570–580, https://doi.org/10.1080/13669877.2018.1443491.

25. Roozenbeek J, Van Der Linden S: **Breaking Harmony Square: a game that "inoculates" against political misinformation**. *Harv. Kennedy Sch. Misinformation Rev.* 2020, https://doi.org/10.37016/mr-2020-47.

26. Goldberg B: **Defanging disinformation's threat to Ukrainian refugees**. *Jigsaw* 2023, **14**. https://medium.com/jigsaw/defanging-disinformations-threat-to-ukrainian-refugees-b164dbbc1c60 (accessed Jul. 18, 2023).

27. Maertens R, Roozenbeek J, Basol M, van der Linden S: **Long-term effectiveness of inoculation against misinformation: three longitudinal experiments**. *J Exp Psychol Appl* Mar. 2021, **27**:1–16, https://doi.org/10.1037/xap0000315.

28. * Rędzio AM, Izydorczak K, Muniak P, Kulesza W, Doliński D: **Is the COVID-19 bad news game good news? Testing whether creating and disseminating fake news about vaccines in a computer game reduces people's belief in anti-vaccine arguments**. *Acta Psychol* Jun. 2023, **236**, 103930, https://doi.org/10.1016/j.actpsy.2023.103930.
The authors tested whether a new gamified inoculation intervention called COVID-19 Bad News, in which players create and spread fake news concerning the COVID-19 pandemic and vaccines, improves people's eagerness to vaccinate. They found that the game did not enhance readiness to vaccinate or reduce believability of unfavourable vaccine-related statements.

29. Basol M, Roozenbeek J, Berriche M, Uenal F, McClanahan WP, van der Linden S: **Towards psychological herd immunity: cross-cultural evidence for two prebunking interventions against COVID-19 misinformation**. *Big Data Soc* Jan. 2021, **8**, 205395172110138, https://doi.org/10.1177/20539517211013868.

30. Iyengar A, Gupta P, Priya N: **Inoculation against conspiracy theories: a consumer side approach to India's fake news problem**. *Appl Cognit Psychol* Sep. 2022, https://doi.org/10.1002/acp.3995. acp.3995.

31. Harjani T, Basol M-S, Roozenbeek J, van der Linden S: **Gamified inoculation against misinformation in India: a randomized control trial**. *J. Trial Error* Feb. 2023, https://doi.org/10.36850/e12.

32. Lyons BA, Montgomery JM, Guess AM, Nyhan B, Reifler J: **Overconfidence in news judgments is associated with false news susceptibility**. *Proc Natl Acad Sci USA* Jun. 2021, **118**, e2019527118, https://doi.org/10.1073/pnas.2019527118.

33. Guess AM, Lerner M, Lyons B, Montgomery JM, Nyhan B, Reifler J, Sircar N: **A digital media literacy intervention increases discernment between mainstream and false news in the United States and India**. *Proc Natl Acad Sci USA*

Jul. 2020, **117**:15536–15545, https://doi.org/10.1073/pnas.1920498117.

34. Roozenbeek J, Traberg CS, Van Der Linden S: **Technique-based inoculation against real-world misinformation**. *R Soc Open Sci* May 2022, **9**, 211719, https://doi.org/10.1098/rsos.211719.

35. Roozenbeek J, van der Linden S: **The fake news game: actively inoculating against the risk of misinformation**. *J Risk Res* 2018, **22**:570–580, https://doi.org/10.1080/13669877.2018.1443491.

36. van der Linden S, Roozenbeek J, Compton J: **Inoculating against fake news about COVID-19**. *Front Psychol* Oct. 2020, **11**, 566790, https://doi.org/10.3389/fpsyg.2020.566790.

37. Graham ME, Skov B, Gilson Z, Heise C, Fallow KM, Mah EY, Lindsay DS: **Mixed news about the bad news game**. *J. Cogn.* 2023, https://doi.org/10.5334/joc.324.

38. Lu C, Hu B, Li Q, Bi C, Ju X-D: **Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: systematic review and meta-analysis**. *J Med Internet Res* Aug. 2023, **25**, e49255, https://doi.org/10.2196/49255.

39. Kellen D, Winiger S, Dunn JC, Singmann H: **Testing the foundations of signal detection theory in recognition memory**. *Psychol Rev* Nov. 2021, **128**:1022–1050, https://doi.org/10.1037/rev0000288.

40. Rathje S, Roozenbeek J, Van Bavel JJ, van der Linden S: **Accuracy and social motivations shape judgements of (mis)information**. *Nat. Hum. Behav.* 2023, https://doi.org/10.1038/s41562-023-01540-w.

41. Modirrousta-Galian A, Higham PA, Seabrooke T: **Wordless wisdom: the dominant role of tacit knowledge in true and fake news discrimination**. *J. Appl. Res. Mem. Cogn.* 2023, https://doi.org/10.1037/mac0000151.

42. Pennycook G, Rand DG: **Lazy, not biased: susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning**. *Cognition* 2019, **188**:39–50, https://doi.org/10.1016/j.cognition.2018.06.011.

43. Bronstein MV, Pennycook G, Bear A, Rand DG, Cannon TD: **Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking**. *J. Appl. Res. Mem. Cogn.* Mar. 2019, **8**:108–117, https://doi.org/10.1037/h0101832.

44. * Nahm FS: **Receiver operating characteristic curve: overview and practical use for clinicians**. *Korean J. Anesthesiol.* Feb. 2022, **75**:25–36, https://doi.org/10.4097/kja.21209.
The author provides an overview of receiver operating characteristic (ROC) curves and their practical significance for clinicians, particularly when determining the presence or absence of diseases. The paper describes the fundamental concepts of ROC curves, as well as their related measures, including the area under the curve (AUC).

45. Pastore RE, Scheirer CJ: **Signal detection theory: considerations for general application**. *Psychol Bull* Dec. 1974, **81**:945–958, https://doi.org/10.1037/h0037357.

46. Pollack I, Hsieh R: **Sampling variability of the area under the ROC-curve and of d'e**. *Psychol Bull* Mar. 1969, **71**:161–173, https://doi.org/10.1037/h0026862.

47. Higham PA, Higham DP: **New improved gamma: enhancing the accuracy of Goodman–Kruskal's gamma using ROC curves**. *Behav Res Methods* Feb. 2019, **51**:108–125, https://doi.org/10.3758/s13428-018-1125-5.

48. Donaldson W, Good C: **A′r : an estimate of area under isosensitivity curves**. *Behav Res Methods Instrum Comput* Dec. 1996, **28**:590–597, https://doi.org/10.3758/BF03200547.

49. Roozenbeek J, Traberg CS, Van Der Linden S: **Correction: "Technique-based inoculation against real-world misinformation" (2023), by Roozenbeek et al**. *R Soc Open Sci* Dec. 2023, **10**, 231235, https://doi.org/10.1098/rsos.231235.

50. Donaldson W: **Measuring recognition memory**. *J Exp Psychol Gen* 1992, **121**:275–277, https://doi.org/10.1037/0096-3445.121.3.275.