Gamified Inoculation Interventions Do Not Improve Discrimination Between True and Fake News: Reanalyzing Existing Research With Receiver Operating Characteristic Analysis

Ariana Modirrousta-Galian & Philip A. Higham

University of Southampton

Word Count: 18,646

Author Note

Ariana Modirrousta-Galian https://orcid.org/0000-0003-2925-2976

Philip A. Higham https://orcid.org/0000-0001-6087-7224

The data and analytic code needed to replicate the analyses reported in this paper and the supplemental materials are available on the Open Science Framework: https://osf.io/85be7/. This study was not preregistered. We have no conflicts of interest to disclose. Our work was funded by the Economic and Social Research Council South Coast Doctoral Training Partnership.

This manuscript was uploaded as a preprint to PsyArXiv (https://psyarxiv.com/4bgkd/), and portions of its findings were presented at the 63rd Annual Meeting of the Psychonomic Society, Boston, MA, United States, November 17–20, 2022.

Ariana Modirrousta-Galian played lead role in data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization and writing of original draft, equal role in writing of review and editing, and supporting role in conceptualization. Philip A. Higham played lead role in conceptualization and supervision, supporting role in data curation, formal analysis, investigation, methodology, project administration, validation, visualization and writing of original draft and equal role in writing of review and editing.

We would like to thank Tina Seabrooke for her proofreading of the manuscript and valuable suggestions.

Correspondence concerning this article should be addressed to Ariana Modirrousta-Galian, Centre for Perception and Cognition, School of Psychology, University of Southampton, Highfield, Southampton, UK, SO17 1BJ. Email: amg1g17@soton.ac.uk

Abstract

Gamified inoculation interventions designed to improve detection of online misinformation are becoming increasingly prevalent. Two of the most notable interventions of this kind are Bad News and Go Viral!. To assess their efficacy, prior research has typically used pre-post designs in which participants rated the reliability or manipulativeness of true and fake news items before and after playing these games, while most of the time also including a control group who played an irrelevant game (Tetris) or did nothing at all. Mean ratings were then compared between pre-tests and post-tests and/or between the control and experimental conditions. Critically, these prior studies have not separated response bias effects (overall tendency to respond “true” or “fake”) from discrimination (ability to distinguish between true and fake news, commonly dubbed discernment). We reanalyzed the results from five prior studies using receiver operating characteristic (ROC) curves, a method common to signal detection theory (SDT) that allows for discrimination to be measured free from response bias. Across the studies, when comparable true and fake news items were used, Bad News and Go Viral! did not improve discrimination, but rather elicited more “false” responses to all news items (more conservative responding). These novel findings suggest that the current gamified inoculation interventions designed to improve fake news detection are not as effective as previously thought and may even be counterproductive. They also demonstrate the usefulness of ROC analysis, a largely unexploited method in this setting, for assessing the effectiveness of any intervention designed to improve fake news detection.

Keywords: Online misinformation, gamification, fake news games, receiver operating characteristics, signal detection theory

Public Significance Statement

This study suggests that Bad News and Go Viral!, two popular online browser games, do not improve people's ability to discriminate between true and fake news. Instead, they cause people to respond more conservatively (i.e., a general tendency to rate all news items as more "false"). This finding highlights the possibility that certain games designed to improve people's ability to spot online misinformation may be counterproductive, as they could be increasing distrust in legitimate information. We offer a key recommendation to avoid this potential risk: Researchers should assess how these gamified psychological interventions affect belief in both true and fake news with a method that can measure discrimination free from response bias, such as receiver operating characteristic analysis.

Gamified Inoculation Interventions Do Not Improve Discrimination Between True and Fake News: Reanalyzing Existing Research With Receiver Operating Characteristic Analysis

"Falsehood flies, and truth comes limping after it, so that when men come to be undeceived, it is too late; the jest is over, and the tale hath had its effect" (Jonathan Swift, 1710/2012, para. 9).

Swift's insight appeared in *The Art of Political Lying*, which was published over 300 years ago. It suggests that attempts to manipulate people with false or misleading information, also known as misinformation, are not novel (Allen et al., 2020). However, recent advances in technology have only served to facilitate the distribution of misinformation. The creation of the internet in 1983 eventually gave rise to *online* misinformation, particularly after the popularization of social media, and provided a global medium for falsehoods as well as the tools necessary to promote their spread (Allen et al., 2020). Seeing as 63% of the total world population uses the internet as of April 2022 (Johnson, 2022), a vast number of people can generate online misinformation that has the potential to reach a huge audience.

Online misinformation, commonly referred to as *fake news*,[1] originates from a variety of sources, including individual social media users, websites, social media influencers, celebrities, and governments (Mukhtar, 2021). The prevalence of misinformation, coupled with people's inability to identify and disregard it, has created a major problem in modern society (Kanozia et al., 2021). As highlighted in Swift's comment, this issue is further exacerbated by the fact that once people have accepted misinformation as true, they often continue to believe in it even after it has been corrected, which is termed the *continued influence effect* (Johnson & Seifert, 1994; Lewandowsky et al., 2012).

[1] Throughout this paper, we use the terms "fake news" and "misinformation" interchangeably for ease of exposition. However, it is important to note that the two terms are not always regarded as synonymous. For example, Lazer et al. (2018) defined fake news as a specific type of misinformation that "mimics news media content in form but … lack[s] the news media's editorial norms and processes for ensuring the accuracy and credibility of information" (p. 1094).

Regardless of its origin, online misinformation can have devastating consequences, such as inciting violence and even threatening democracy. This is exemplified by the deadly insurrection at the US Capitol on January 6, 2021, which was fueled by false election fraud claims that proliferated on social media at the time (Calvillo et al., 2021). Considering the alarming effects of online misinformation, finding ways to combat this issue is imperative. For psychologists, an important avenue of investigation involves finding ways to improve people's ability to identify and thus discount online misinformation before it is accepted as true, thereby avoiding the problem of continued influence. Early misinformation discounting limits the unintentional spread of online misinformation by social media users (Adjin-Tettey, 2022), reduces the incentive to create false or misleading content due to a lack of public confidence and support (Van Bavel et al., 2021), and prevents the harmful consequences that stem from mistakenly believing online misinformation (Ecker et al., 2022).

Inoculation Theory and Gamified Interventions

Gamified psychological interventions designed to protect people from online misinformation before it is encountered are becoming increasingly prevalent. Gamification refers to using game design elements in non-game settings (Huotari & Hamari, 2016), and its popularity in the context of online misinformation can be attributed to its ability to increase public participation and stimulate user engagement (Morschheuser et al., 2016). The best-known gamified fake news interventions are typically informed by inoculation theory (McGuire, 1964). Inoculation theory draws upon a medical analogy; vaccines containing a weakened dose of a virus can trigger the production of antibodies in the immune system that confer resistance against future infection by a stronger version of the same virus. Analogously, inoculation theory suggests that by exposing people to weakened arguments against an attitude they hold, resistance can be conferred against future attacks on that particular attitude (Banas & Rains, 2010). In recent years, inoculation theory has been applied to the topic of online misinformation, with researchers investigating the possibility of creating a "broad-spectrum vaccine" that generates "mental antibodies" against false or misleading content spread on the internet (Roozenbeek & van der Linden, 2019, p. 2). To

achieve psychological inoculation, gamified interventions have exposed people to some common techniques used to produce online misinformation. The hope is that such exposure will confer resistance against actual online misinformation they encounter in the future (Roozenbeek & van der Linden, 2019).

Gamified interventions that aim to pre-emptively protect people against online misinformation utilize “prebunking”, namely, preventing people from believing online misinformation they encounter in the future (Tay et al., 2021). Prebunking differs from the more traditional “debunking”, which aims to retrospectively correct people’s belief in online misinformation. Although the comparative effectiveness of prebunking and debunking falls outside the scope of this paper, it is worth noting that the research comparing the efficacy of these approaches is mixed; some studies show that debunking is more effective at improving news veracity discernment than prebunking (Brashier et al., 2021), while others indicate the opposite (Grady et al., 2021). Nevertheless, interventions derived from inoculation theory automatically incorporate prebunking due to their preventative nature.

Intervention Specificity

The overarching aim of any psychological intervention is to change people’s behavior for the better (Jhangiani et al., 2019). However, some interventions can be over-general and unintentionally influence other behaviors. Depending on the behaviors, overly general interventions can be devastating. For example, in the last few decades, interventions have been developed to reduce the blame associated with mental illness. This blame reduction has been achieved by developing interventions that place more emphasis on the biological causes of mental illness (Corrigan, 2016). Phelan et al. (2011) found that these interventions helped people understand that mental illness is not a choice, but more akin to a disease. When only taking this outcome into account, this type of intervention may seem successful, but it is only half the story. Unfortunately, these interventions also promoted the belief that mental illness is hard wired and untreatable, which can influence whether an employer will hire people with mental illness, or whether a landlord will rent to them. Thus, despite the interventions having the desirable intended effect, the unfortunate unintended effects have

meant that the success of these interventions have been called into question (Corrigan, 2016; Phelan et al., 2011; Read & Harré, 2009; Read, 2011).

In our view, the issue of online misinformation is similar to this example from clinical psychology. If an intervention has the desired effect of decreasing belief in fake news but also has the undesired effect of decreasing belief in true news, then we should question its efficacy. As a case in point, Clayton et al. (2020) found that when people received a general warning about misleading information on social media, their belief in both true and false headlines decreased. Accordingly, they concluded that these general warnings "pose a potential hazard" as they could "increase distrust in legitimate information" (p. 1091). In fact, failure to believe the truth can potentially be more damaging than believing falsehoods. For example, rejecting the valid scientific evidence supporting the efficacy of COVID-19 vaccines is arguably more harmful than believing the falsehood that 5G towers cause COVID-19. The former leads to vaccine refusal, which threatens personal and global health, while the latter leads to vandalization of 5G towers, which is comparably innocuous (Afolabi & Ilesanmi, 2021; Pertwee et al., 2022). Therefore, in our view, any intervention designed to tackle the issue of online misinformation must be assessed both with respect to belief in fake news *and* belief in true news (see also Guay et al., 2022). However, as will be demonstrated in subsequent sections, this is not a universal opinion in the field of online misinformation research.

Signal Detection Theory and Receiver Operating Characteristic Analysis

Batailler et al. (2022) were the first to apply signal detection theory (SDT) to research on the identification of fake news. They used SDT to reanalyze data from two previous studies: Pennycook et al. (2018) and Pennycook and Rand (2019). The reanalyses provided more nuanced insights into the results of these studies by distinguishing between discrimination and response bias. In the context of online misinformation, discrimination refers to the ability to distinguish between true and fake news, which is also referred to as *news veracity discernment*, while response bias refers to the general tendency to rate news items as true or fake regardless of their objective veracity. If an intervention affects only the

target behavior, this will result in an increase in discrimination. For example, discrimination would improve if belief in fake news decreases while belief in true news remains relatively intact. Conversely, if the intervention has more general effects on both true and fake news by, for example, reducing belief in all news regardless of objective veracity, this will affect response bias.[2]

Batailler et al. (2022) used single-point indices of discrimination (*d'*) and response bias (*c*), which carry the strong assumption of equal-variance Gaussian underlying evidence distributions. Receiver operating characteristic (ROC) analysis, a more powerful methodology common to SDT, allows researchers to measure discrimination free from response bias without making the same strong assumptions (Higham & Higham, 2018). Indeed, Batailler et al. recommended using ROC analysis in future research. To the best of our knowledge, ROC analysis has only been used once before with research on fake news (Modirrousta-Galian et al., in press). Instead, most studies have analyzed mean belief ratings in true and fake news, which are not ideally suited for separating discrimination and response bias.

SDT assumes that people have an internal dimension, also referred to as a decision axis, that represents the amount of subjective evidence for the presence of one type of stimulus over another type of stimulus (Aleci, 2021). Commonly, one type of stimulus contains sought-after information (e.g., truth) and is called a *signal* trial, whereas this information is absent in the other type of stimulus and is called a *noise* trial.[3] To understand this, consider the discernment of true and fake news items. When the task is defined as detecting truth in news items, the internal dimension will represent a continuum ranging from

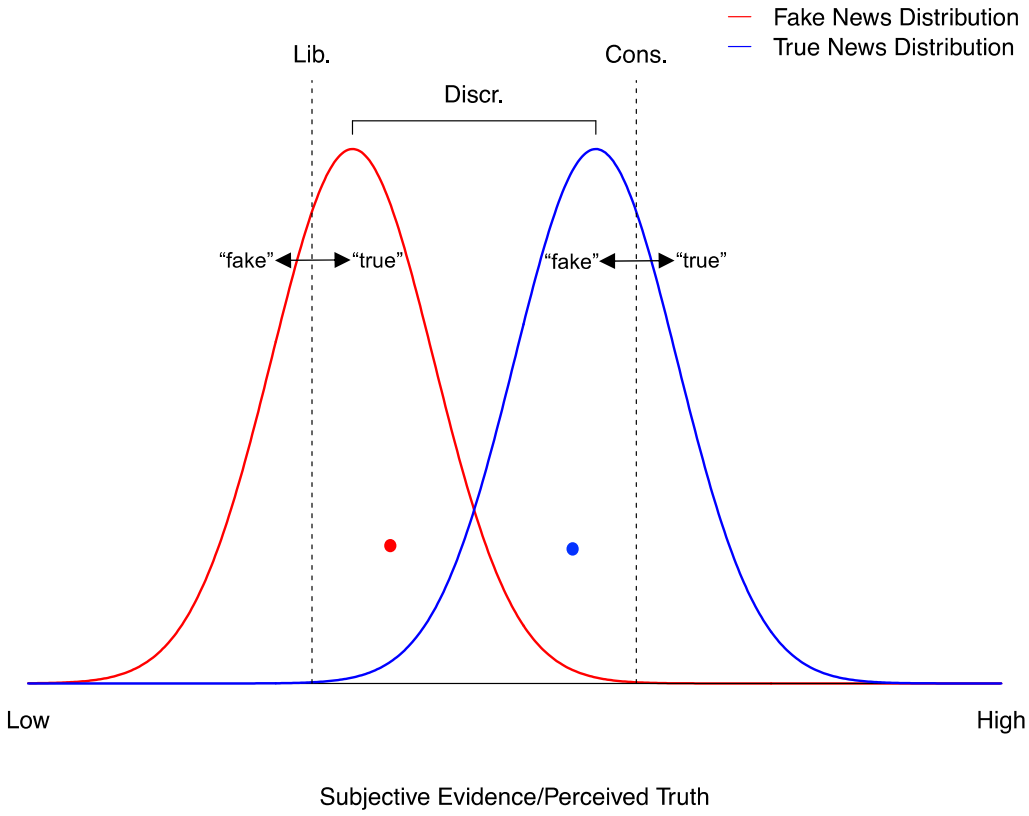
[2] A change in response bias is often interpreted to mean that the placement of the response criterion (the decision cut-off, as explained later) has been altered. However, although a change to the placement of the response criterion will affect measures of response bias, so will changes to the placement of the evidence distributions on the subjective evidence dimension. We elaborate on these different interpretations of response bias effects in the General Discussion. For now, readers should not assume that if we describe an intervention as affecting response bias, we do not necessarily mean that the response criterion has moved.

[3] SDT applications are commonly concerned with modelling discrimination when there is one type of evidence associated with the signal stimulus (e.g., evidence for truth) which is lacking from the noise stimulus. However, it is possible that noise trials could contain evidence aiding discrimination as well. For example, a fake news item (noise) might not just lack evidence for truth, it might also contain evidence for falsity.

low perceived truth to high perceived truth (see Figure 1). The subjective mapping of true and fake news items on this internal dimension can be represented by equal-variance Gaussian distributions as shown in Figure 1, although other possibilities can also be considered (e.g., unequal-variance Gaussian distributions). Gaussian distributions are commonly assumed because of noise; that is, the perceived truth of a news item, regardless of whether it is true or fake, will vary across items and be influenced by other random factors such as memory and current context (Heeger, 1997; Higham et al., 2016). Notably, since the task is to detect truth in news items, the true news (signal) distribution will be higher on the internal dimension (i.e., further to the right) than the fake news (noise) distribution, as long as discernment is above chance.

Figure 1

Equal-Variance Gaussian Distributions of True and Fake News Items Distributed Over a Dimension of Subjective Truth as Conceptualized by Signal Detection Theory



Note. Lib. = Liberal; Cons. = Conservative; Discr. = Discrimination. The vertical dotted lines represent liberal (left) and conservative (right) decision criteria. The red and blue circles represent individual fake and true news items, respectively.

Figure 1 shows the simplest SDT case with two types of stimuli (true and fake news) and two available responses ("true" and "fake"). The response elicited for a given item is determined by the item's position relative to a cut-off, also known as a response criterion. The placement of the criterion on the decision axis is malleable and assumed to be under the control of the observer (Aleci, 2021). We have represented this malleability by including two criteria (vertical dotted lines) in two different positions in Figure 1. If the subjective evidence of an item is equal to or higher than the criterion, the observer gives a response indicating the presence of a signal, which in this case would be a "true" response. If the

subjective evidence associated with an item does not reach the criterion, then the observer responds that the signal is absent (i.e., “fake”). If the criterion is placed low on the internal dimension (e.g., the vertical line further to the left in Figure 1), then little subjective evidence is needed for a “true” response, so the observer responds “true” frequently. In this case, the observer is said to have a liberal response bias. In contrast, if the criterion is placed high on the internal dimension (e.g., the vertical line further to the right in Figure 1), then a considerable amount of subjective evidence is needed for a “true” response, so the observer responds “true” relatively infrequently. In this case, the observer is said to have a conservative response bias. Finally, the ability for the observer to accurately discriminate between true and fake news items is represented by the overlap of the distributions (i.e., complete overlap = no discrimination; no overlap = excellent discrimination). A common measure of discrimination is the standardized distance between the means of the respective distributions, which is known as *d'* (de Gardelle & Kouider, 2009). Note that discrimination is the same regardless of whether the response criterion is liberal or conservative in Figure 1, reflecting the independence of discrimination and response bias.

Also shown in Figure 1 are two specific news items represented as circles, one fake (red) and the other true (blue). For the liberal criterion case, the evidence associated with both items exceeds the criterion, so the observer would respond “true” for both. This is a correct response for the true (blue) news item and is called a *hit.* The proportion of all true news items falling above the criterion is called the *hit rate (HR),* which for the liberal case in Figure 1 is approaching 1.0. The “true” response to the fake (red) news item, however, is a type of error called a *false alarm*. The proportion of fake news items falling above the criterion (which in this case is liberal) is called the *false alarm rate* (*FAR*), which in the example in Figure 1 is about 0.7.

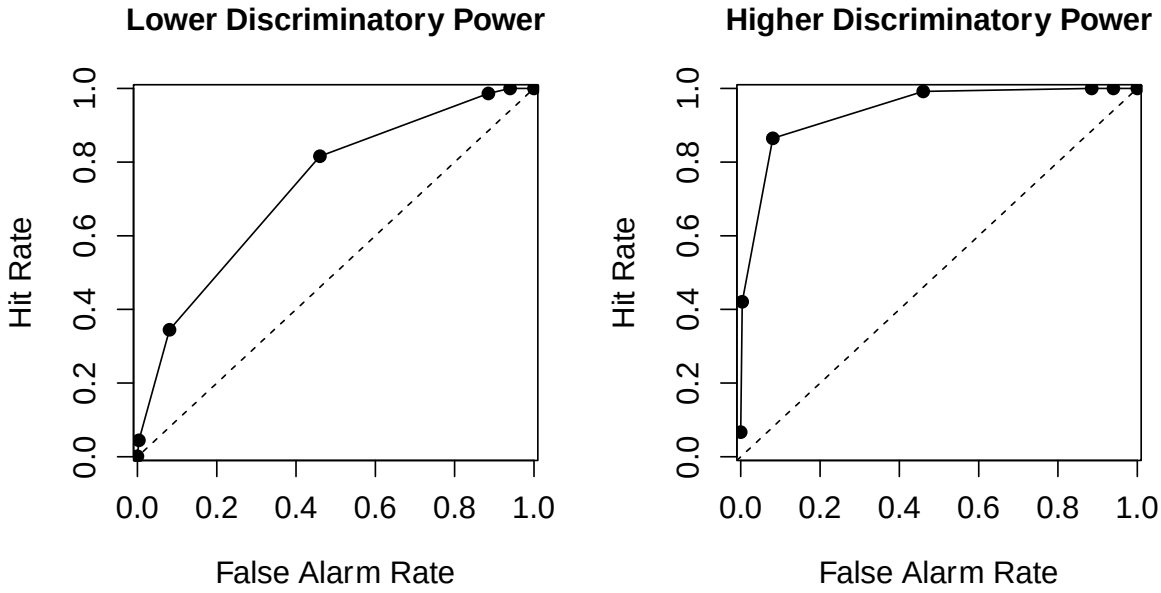
The situation is somewhat different for the conservative criterion case. Now, neither item exceeds the criterion, and so the observer responds “fake” to both. This response is correct for the fake news item and is called a *correct rejection*. The *correct rejection rate* (i.e., the proportion of correct rejections) is equal to one minus the FAR. The “fake” response

to the true news item is incorrect and is called a *miss*. The *miss rate* (i.e., the proportion of misses) is equal to one minus the HR.

Figure 1 depicts the SDT model for binary tasks, where participants can only choose between two answers (e.g., “true” or “fake”). If participants answer using an ordinal rating scale, however, a more powerful ROC analysis can be conducted instead. Two example ROC curves are shown in Figure 2. In short, ROC curves plot multiple HRs as a function of multiple FARs, which are each derived from the individual points on the rating scale. ROC curves provide a useful graphical tool for visualizing discrimination and response bias (Tasche, 2008). The ordinal rating scale is dichotomized at each point on the scale by treating each level (e.g., 1, 2, 3, 4, 5, 6, and 7) as a single cut-off point, specifically a point on the scale that corresponds to hypothetical yes/no (or in this case true/fake) criteria (Mandrekar, 2010). A HR and a FAR is computed for each scale value, with higher scale values corresponding to more conservative criteria. So, if 4 was the cut-off point, the proportion of true and fake news items assigned 4 or higher would constitute the HR and FAR, respectively. This process is completed for all the points on a scale. Chance-level discrimination, where the HR equals the FAR for all points (i.e., complete overlap of the evidence distributions), corresponds to a straight line drawn from the bottom-left to the top-right in the ROC space. It is commonly included as a point of reference when plotting ROC curves (see Figures 2 & 3).

Figure 2

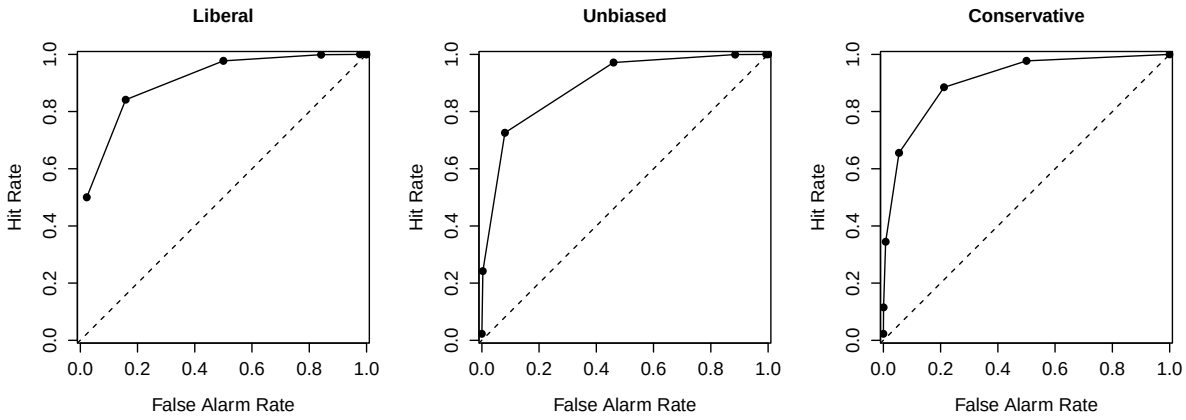
ROC Curves Showing Lower Discriminatory Power and Higher Discriminatory Power



Note. ROC = Receiver operating characteristic. These ROC curves are obtained from equal-variance Gaussian distributions.

Figure 3

ROC Curves Showing a Liberal Response Bias, no Response Bias, and a Conservative Response Bias



Note. ROC = Receiver operating characteristic. These ROC curves are obtained from equal-variance Gaussian distributions, and discrimination was kept constant across all three scenarios. Although it appears as though the liberal figure has five points and the unbiased figure has six points, in fact all figures have seven points; three points in the upper-right portion of the liberal figure are overlapping, and two points in the upper-right portion of the unbiased figure are overlapping.

ROC curves can be created for individual participants by plotting their HRs and FARs at each scale point. Alternatively, ROC curves can be generated for different tests (e.g., pre-tests and post-tests) or for different experimental conditions (e.g., treatment conditions and control conditions) by plotting HRs and FARs at each scale point aggregated across participants. The shape of the ROC curve and where the points lie on it provide a straightforward means to examine discrimination and response bias. Changes in discrimination are indicated by the extent to which the ROC curve bows from the diagonal; if discrimination decreases, the ROC curve bows less from the diagonal, whereas if discrimination increases, the ROC curve bows further from the diagonal (see Figure 2). In contrast, changes in response bias are indicated by the placement of the points on the ROC curve; a conservative response bias is indicated by ROC points that are more offset towards

the lower-left portion of the curve, whereas a liberal response bias is indicated by ROC points that are more offset towards the upper-right portion of the curve (see Figure 3).

Statistical measures of discrimination and response bias can also be obtained from ROC curves. One of the most widely used non-parametric measures of discrimination is the area under the curve (AUC). As discrimination increases from chance-level responding to perfect discrimination, AUC increases from .5 to 1.0. It can be estimated by using the trapezoidal rule formula from Pollack and Hsieh (1969):

$$AUC = 0.5 \sum_{k=0}^{n} (HR_{k+1} + HR_k)(FAR_{k+1} - FAR_k) \quad (1)$$

In this equation, *k* denotes the different criteria plotted on the ROC curve and *n* represents the total number of criteria. The trapezoidal rule estimates the AUC by drawing straight lines between the points on the ROC curve to create trapezoids. Therefore, if the ROC curve is curvilinear, the trapezoidal rule underestimates the AUC (DeLong et al., 1988). Corrections have been offered to compensate for this underestimation (see Donaldson & Good, 1996), but they are mathematically complex and limited to certain types of data (Higham & Higham, 2018). Therefore, we used the trapezoidal rule in the current paper as it has the advantage of broad applicability, few assumptions, and mathematical simplicity (Messori et al., 2019).

A useful non-parametric index of response bias is *B"D*, which can be calculated at each scale point using the following formula from Donaldson (1992):

$$B"D = \frac{(1 - HR)(1 - FA) - (HR)(FA)}{(1 - HR)(1 - FA) + (HR)(FA)} \quad (2)$$

Crucially, its usefulness stems from its ability to provide accurate estimates of response bias even when it is calculated from collapsed or grouped data, and also over the full range of discrimination performance, namely from chance to perfect performance (Donaldson, 1992; See et al., 1997; Snodgrass & Corwin, 1988). Positive versus negative values of *B"D* correspond to conservative versus liberal responding, respectively, with *B"D* = 0 indicating no response bias and the maximum absolute value being 1.0. *B"D* differs from AUC in that it is computed for a single HR and FAR rather than being based on the whole ROC curve with

multiple points. However, as will become clear, it is possible to compute *B"D* separately for each point on the ROC curve to assess response bias. It should be noted that non-parametric measures of discrimination and response bias, such as the AUC using the trapezoidal rule and *B"D*, have the clear advantage of not making strong assumptions about the nature of the underlying evidence distributions, which may not be met. For example, *d',* which assumes equal-variance Gaussian distributions, is not independent of response bias if the variances of the Gaussian evidence distributions are not equal.

Aims of the Present Paper

As noted earlier, despite its usefulness for assessing the specificity of an intervention's effect, we are aware of only one paper that has analyzed data from gamified fake news interventions using ROC analysis (Modirrousta-Galian et al., in press). Indeed, in some prior research, there is seemingly a lack of concern over the generality of the intervention's effect. For example, some researchers have reported that they collected participants' ratings of true news items, but either excluded them from the main analysis (Maertens et al., 2021) or failed to report them in their manuscripts at all (Basol et al., 2020).

Consequently, the aim of the current paper was to reanalyze data from published papers on gamified fake news interventions using ROC analysis to determine the specificity of the effects. This paper is not intended to be a critique of prior studies, and thus does not discuss their potential methodological problems unless they are directly relevant to the reanalysis. The following criteria had to be met for a study to be included in our reanalysis: (a) the study examined the effectiveness of a gamified fake news intervention; (b) a scale pertaining to news veracity was included as a dependent variable; (c) ratings were collected for both true and fake news items; and (d) the raw data from the study were either publicly available or made available by the authors.

We conducted our literature search in March 2022. First, we searched Google Scholar and ProQuest with the following search query: *("fake news game" OR "fake news intervention" OR "misinformation game" OR "misinformation intervention") AND ("true news items" OR "real news items" OR "fake news items" OR "false news items")*, which revealed

37 potentially eligible studies. We examined the titles, abstracts, and results of these 37 papers and found that five met our inclusion criteria (Basol et al., 2020; Basol et al., 2021; Maertens et al., 2021; Roozenbeek et al., 2020; Roozenbeek & van der Linden, 2019). We then manually searched the reference lists in these five studies and found another potentially eligible paper (Saleh et al., 2021). Finally, we searched Google with the following search query: *"fake news game"* and found three more potentially eligible studies (Grace & Hone, 2019; Micallef et al., 2021; Urban et al., 2019). However, after examining the titles, abstracts, and results of these four additional papers, we found that they did not meet our inclusion criteria. Therefore, out of 41 potential papers, five were selected.

Some of these five papers reported several experiments, and in some cases, it was not possible to reanalyze data from every experiment. Details about which experiments were included or excluded from each of the five studies are provided in more detail later when we describe the specifics of each reanalysis. The five suitable papers all used the games Bad News or Go Viral!. Both fake news games are considered prebunking interventions as they aim to pre-emptively protect people against online misinformation by exposing them to the common techniques used in its production. These interventions have received much attention both in the academic literature and beyond. Bad News alone has been played over a million times and Go Viral! is supported by the UK Cabinet Office, the World Health Organization, and the United Nations. Furthermore, both interventions have been reported on by various mainstream media outlets, such as the BBC and CNN, and their popularity has led to the current (as of December 2022) versions of Bad News and Go Viral! being playable in 19 and 13 different languages, respectively. Considering the impact of these games and the fact that all our reanalyses are based on them, we will now describe them in some detail.

Bad News and Go Viral!

Bad News has players adopt the role of a fake news creator whose aim is to gather as many followers as possible whilst also maintaining credibility. To do this, players use six different strategies to create online misinformation, namely, impersonating people, emotional

language, group polarization, conspiracy theories, discrediting opponents, and trolling. After using each strategy, players receive a fake news badge, which essentially serves as a progression milestone. Players can choose between different options in the game. Most importantly, they can choose from several different fake news stories to share. Players must pay attention to their follower and credibility meters, which are contingent on their choices. If the credibility meter drops to zero, the game ends and the player loses. However, if the credibility meter remains above zero and players use all six strategies to create online misinformation, the game ends and the player wins, and the total number of followers they gathered counts as their final score.

The gameplay of Go Viral! is almost identical to that of Bad News. The main gameplay differences between them are the following: (a) the follower meter in Bad News is replaced with a “likes” meter in Go Viral!; (b) three different strategies are used to create online misinformation in Go Viral!, namely fearmongering, using fake experts, and conspiracy theories, instead of the six used in Bad News; (c) unlike Bad News, players do not receive fake news badges in Go Viral!; and (d) Bad News presents players with general online misinformation, whereas Go Viral! only presents players with COVID-19-related online misinformation. Due to these gameplay differences, particularly the disparity in the number of strategies and the inclusion versus exclusion of fake news badges, Bad News takes about 15 minutes to complete, while Go Viral! takes about 5 minutes.

To determine the effectiveness of Bad News and Go Viral!, prior research has used pre-post designs in which participants rated true and fake news items before and after playing either Bad News (Basol et al., 2020; Maertens et al., 2021; Roozenbeek et al., 2020; Roozenbeek & van der Linden, 2019) or Go Viral! (Basol et al., 2021). All but one of these studies also included a control group that completed ratings before and after either playing Tetris, which is not designed to improve detection of fake news, or not playing anything at all. Studies investigating the effectiveness of Bad News used a 7-point scale ranging from 1 (*not at all reliable*) to 7 (*very reliable*), while the study investigating the effectiveness of Go Viral! used a 7-point scale ranging from 1 (*not at all manipulative*) to 7 (*very manipulative*).

Notably, we will use mean ratings as an umbrella term to refer to both mean reliability ratings and mean manipulativeness ratings throughout this paper.

Mean ratings for fake news items were compared between pre-tests and post-tests, as were mean ratings for true news items in most studies. The results from most of this prior work showed that Bad News and Go Viral! reduced people's belief in fake news, but also sometimes in true news, albeit to a lesser degree. This raises the concern that the effects of Bad News and Go Viral! might be overly general, affecting responding to all news items regardless of their objective veracity. However, Basol et al. (2021) concluded that these types of gamified psychological interventions are effective in helping people detect and discount fake news, and although they also sometimes reduce belief in true news, the variability and small size of this effect suggests that this may be due to item effects rather than general skepticism.

Overview of the Reanalyses

As noted earlier, we reanalyzed the results from five different studies, four of which investigated the effectiveness of Bad News (Basol et al., 2020; Maertens et al., 2021; Roozenbeek et al., 2020; Roozenbeek & van der Linden, 2019) and one of which investigated the effectiveness of Go Viral! (Basol et al., 2021). To do this, we used ROC analysis to assess discrimination free from response bias and thus determine whether the findings from these previous experiments were due to improvements in the ability to discriminate between true and fake news (i.e., the intervention had an effect specific to belief in fake news), shifts in response bias (i.e., the intervention had a general effect on belief in all types of news), or both.

In each reanalysis, we created ROC curves for different tests (i.e., pre-tests and post-tests) and experimental conditions (i.e., treatment conditions and control conditions) aggregated over participants. For statistical analysis, the trapezoidal rule was used to calculate the AUC for each participant, and the *B"D* was calculated at each scale point

(except for 1)[4] for each participant, resulting in six *B"D* values per participant.[5] The *B"D* values were then collapsed across all scale points for each participant to result in one average *B"D* value per participant. ANOVAs and *t*-tests were carried out to compare participants' AUC and *B"D* values between tests (i.e., pre-tests and post-tests). Furthermore, the Bayes factor was calculated for these analyses and interpreted through the discrete evidence categories proposed by Jeffreys (1961) and their corresponding interpretations adapted by Lee and Wagenmakers (2013; see Table 1).

Table 1

Bayes Factor Evidence Categories According to Jeffreys (1961) and Their Corresponding Interpretations Adapted by Lee and Wagenmakers (2013)

BF_{10}	Interpretation
>100	Extreme evidence for H1
30–100	Very strong evidence for H1
10–30	Strong evidence for H1
3–10	Moderate evidence for H1
1–3	Anecdotal evidence for H1
1	No evidence
1/3–1	Anecdotal evidence for H0
1/10–1/3	Moderate evidence for H0
1/30–1/10	Strong evidence for H0
1/100–1/30	Very strong evidence for H0
<1/100	Extreme evidence for H0

Note. BF_{10} quantifies the empirical evidence in favor of the alternative hypothesis.

[4] The *B"D* was not calculated at scale point 1 because both the HR and FAR are necessarily equal to 1.0 due to the nature of the task (i.e., all items are assigned 1 or higher). Hence, it provides no meaningful information.

[5] When participants have a HR of 1 and a FAR of 0, the formula for *B"D* results in 0/0, which is undefined. Therefore, we applied a loglinear correction when calculating the HRs and FARs (only for calculating *B"D*) by adding 0.5 to both the number of hits and false alarms and adding 1 to both the number of signal (true news) and noise (fake news) trials (Stanislaw & Todorov, 1999).

Overview of the Results

A summary of the results are shown in Table 2.

Table 2

Summary of the Results

Study	*n*	AUC analysis					*B"D* analysis				
		$M_{pretest}$	$M_{posttest}$	*p*	*d*	BF_{10}	$M_{pretest}$	$M_{posttest}$	*p*	*d*	BF_{10}
Roozenbeek and van der Linden (2019)	13,564	.88	.91	< .001	0.17	1.73×10^{122}	-.09	.04	< .001	0.40	1.34×10^{486}
Basol et al. (2020), Treatment Condition	96	.73	.75	.359	0.09	0.17	.07	.24	< .001	0.47	48,915.96
Basol et al. (2020), Control Condition	102	.70	.71	.543	0.05	0.13	.03	.07	.090	0.12	0.45
Roozenbeek et al. (2020), Set A–A, Experiment 1	480	.75	.78	.074	0.16	0.48	.21	.36	< .001	0.39	764.51
Roozenbeek et al. (2020), Set B–B, Experiment 1	480	.83	.84	.452	0.07	0.13	.14	.17	.426	0.07	0.14
Roozenbeek et al. (2020), Control Condition, Experiment 2	760	.80	.79	.037	-0.07	0.36	.06	.08	.094	0.05	0.16
Maertens et al. (2021), Treatment Condition, Experiment 1	58	.88	.88	.943	0.12	0.01	-.16	.14	< .001	1.40	5.32×10^{15}
Maertens et al. (2021), Control Condition, Experiment 1	60	.87	.86	.534	-0.20	0.03	-.21	-.14	.002	0.55	7.81
Maertens et al. (2021), Treatment Condition, Experiment 2	54	.74	.77	.482	0.20	0.12	.05	.20	< .001	0.87	289.85
Maertens et al. (2021), Control Condition, Experiment 2	56	.71	.71	.909	0.08	0.06	-.01	.03	.541	0.20	0.10
Basol et al. (2021), Study 1	1,771	.89	.90	.003	0.06	2.48	.03	.13	< .001	0.27	3.07×10^{37}
Basol et al. (2021), Active Condition, Study 2	151	.86	.89	< .001	0.59	4982.90	.05	.21	< .001	0.84	213,083,135
Basol et al. (2021), Control Condition, Study 2	235	.84	.86	< .001	0.46	865.87	.08	.11	.111	0.19	0.13

Note. For Maertens et al. (2021) and Basol et al.'s (2021) Study 2, *d* was converted from η^2 on https://www.psychometrica.de/effect_size.html. $M_{posttest}$ is the average AUC and *B"D* across all post-tests.

Reanalysis of Bad News

Transparency and Openness

The data and analytic code needed to replicate these reanalyses are available on the Open Science Framework: https://osf.io/85be7/. We obtained ethical approval to conduct this research from the University of Southampton Faculty of Environmental and Life Sciences Ethics Committee (77386). This study was not preregistered.

Roozenbeek and van der Linden (2019)

The aim of Roozenbeek and van der Linden's (2019) study was to investigate the effect of Bad News on people's ability to identify misinformation. To test this, they embedded a voluntary pre-post survey in the game that asked players who opted in to rate the reliability of the same five news items before and after playing Bad News. Reliability ratings were made on a scale that ranged from 1 (*not at all reliable*) to 7 (*very reliable*), and the five news items were presented in the form of Twitter posts and news headlines. Three of these news items were created by the researchers and contained false information, and two were obtained from global news events and contained true information. The three fake news items reflected three out of the six misinformation techniques presented in Bad News (one technique per item), namely impersonating people, floating conspiracy theories, and discrediting opponents.

Roozenbeek and van der Linden (2019) hypothesized that participants would rate the fake news items, but not the real news items, as less reliable after playing Bad News compared to before. To test this hypothesis, differences in mean reliability ratings between the pre-test and the post-test were analyzed for each of the five news items. The results showed that although the pre-post differences were statistically significant for all items, the effect sizes for the true news items were almost negligible (i.e., $d \leq 0.04$), whereas the effect sizes for the fake news items were much greater (i.e., $d \geq 0.30$). Given the large sample size (14,163–14,266 depending on the news item), Roozenbeek and van der Linden concluded that there were meaningful pre-post differences for the fake news items but not for the real news items, which supported their hypothesis.

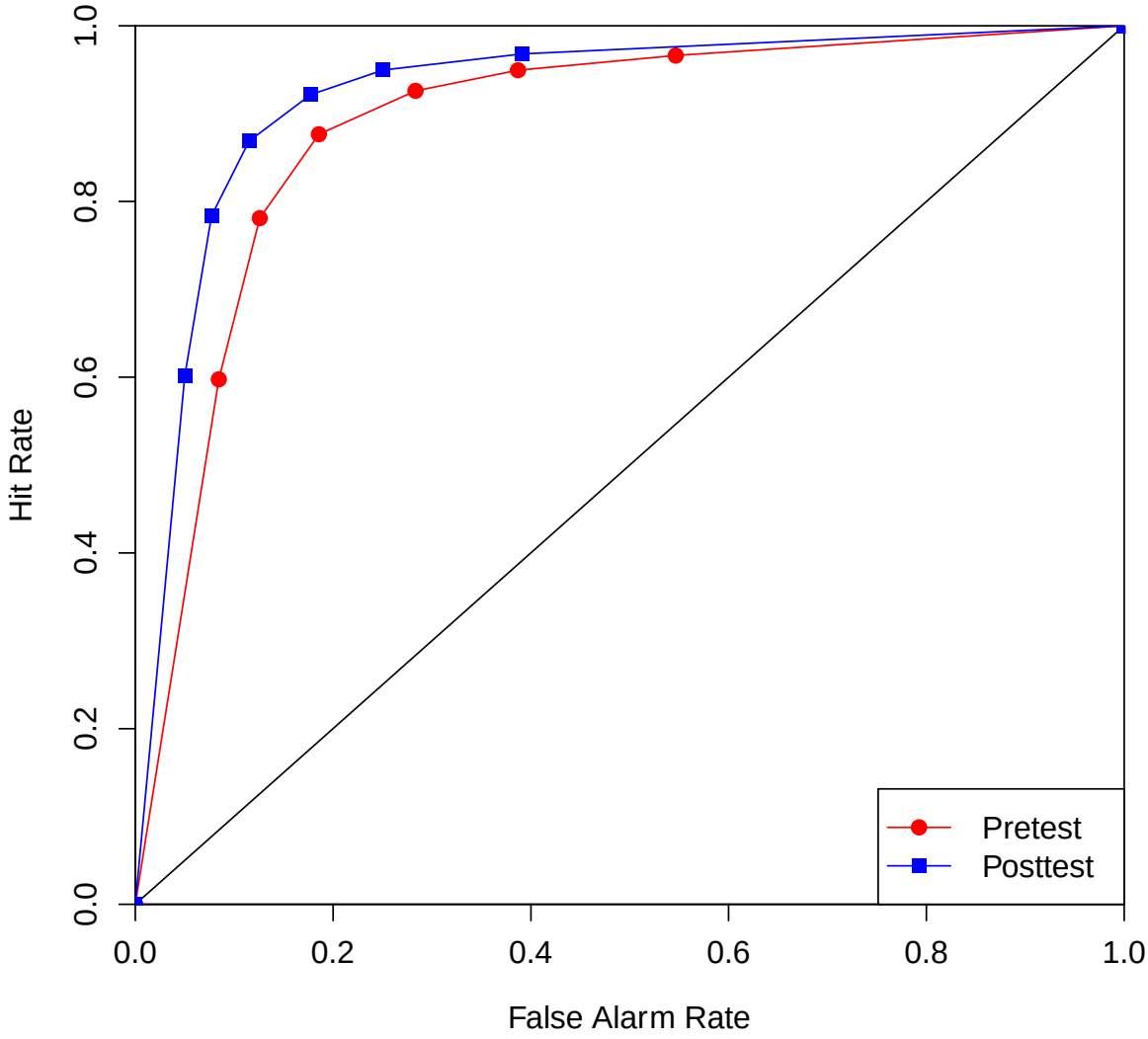
It should be noted that Roozenbeek and van der Linden's (2019) study involved two separate data collection stages: The first is described above, and the second gathered and analyzed pre-post data for an additional fake news item. This item was also created by the researchers and presented in the form of a news headline, but it reflected a different misinformation technique, namely, polarizing opponents. A different, much smaller set of participants took part in the second data collection stage (885) compared to the first data collection stage (14,163–14,266). The second data collection stage only gathered pre-post data for one fake news item and no true news items. Omitting the true news items made it impossible to compute HRs, and consequently AUC and *B"D* values. Therefore, we limited our reanalysis to the initial data collection stage and to participants who had completed the pre-test and the post-test for all five news items. This resulted in a sample of 13,564 participants for our reanalysis.

Results

The ROC curves for the pre-test and the post-test are shown in Figure 4. A paired samples *t*-test revealed that although the AUC values for the pre-test (M = .88, SD = .19) were significantly smaller than the AUC values for the post-test (M = .91, SD = .17), the effect size was almost negligible, $t(13{,}563) = 24.18$, $p < .001$, $d = 0.17$, 95% CI [0.16, 0.18]. However, the Bayes factor indicated extreme evidence in favor of the alternative hypothesis, $BF_{10} = 1.73 \times 10^{122}$. Although the HRs remained mostly equal across the pre-test and the post-test, the FARs decreased (see Table S1). A paired samples *t*-test revealed that the *B"D* values for the pre-test (M = -.09, SD = .55) were significantly smaller than the *B"D* values for the post-test (M = .04, SD = .51), $t(13{,}563) = 49.45$, $p < .001$, $d = 0.40$, 95% CI [0.38, 0.41] (see Table S2), and the Bayes factor indicated extreme evidence in favor of the alternative hypothesis, $BF_{10} = 1.34 \times 10^{486}$.

Figure 4

ROC Curves for the Pre-Test and Post-Test in Roozenbeek and van der Linden (2019)



Note. ROC = Receiver operating characteristic.

Discussion

In summary, Bad News improved participants' news veracity discernment, although the effect size was almost negligible. Specifically, the FAR for fake news decreased, whereas the HR for true news was unaffected, a result that is consistent with larger pre-post differences in reliability ratings for the fake news items compared to the true news items reported in Roozenbeek and van der Linden (2019). This pattern of responding resulted in

higher *B"D* values because the overall proportion of responses exceeding the upper criteria (i.e., those associated with scale values greater than 1) decreased (cf. Tables S1 and S2).

On the surface, this outcome suggests that Bad News is a mildly effective intervention that targets belief in fake news while having no effect on true news. However, before accepting this interpretation out of hand, it is worth taking a closer look at the items that were used in this study. Specifically, the two true news items were evidently reliable, having been reported extensively in the mainstream media (i.e., “President Trump wants to build a wall between the United States and Mexico” and “#Brexit, the United Kingdom’s exit from the European Union, will officially happen in 2019”). This obvious reliability explains why the mean pre-test reliability rating across the two true news items was 6.10, which is almost at the upper limit of the scale (i.e., 7). In contrast, the fake news items were ambiguous, having been created by the researchers (e.g., “The 8th season of #GameOfThrones will be postponed due to a salary dispute”). This ambiguity resulted in an average pre-test reliability rating across the three fake news items of 2.61, which is comparatively closer to the mid-point of the scale (i.e., 4).

It is unlikely that any psychological intervention would impact belief in true news items that are near the ceiling of the reliability scale because they have been reported extensively in the mainstream media. Participants’ memories of those reports would likely immunize their ratings to any sort of change. Hence, it may be premature to conclude that Bad News is inherently an intervention that targets only fake news. It is equally plausible that its effects are general, but the particular true news items used in this study masked this generality. Fortunately, other research investigating the efficacy of Bad News has used true news items that are less obviously true and more comparable to the fake news items, allowing for some level of uncertainty. This research provides a better test of the specificity of Bad News’ effects, and we turn to that research next.

Basol et al. (2020)

The purpose of Basol et al.’s (2020) study was to replicate the findings reported in Roozenbeek and van der Linden (2019) with a more robust experimental design.

Specifically, they added a randomized control condition in which participants played Tetris instead of Bad News. Additionally, they used a larger set of news items to test participants' ability to spot misinformation. Both conditions included a pre-test and a post-test that asked participants to rate the reliability of the same 21 news items before and after playing either Tetris or Bad News. Reliability ratings were made on a scale that ranged from 1 (*not at all reliable*) to 7 (*very reliable*), and all news items were presented in the form of Twitter posts. Eighteen of these news items were created by the researchers and contained false information, and three were obtained from global news events and contained true information. There were three fake news items corresponding to each of the six misinformation techniques presented in Bad News.

To assess the effectiveness of Bad News, differences in mean reliability ratings between the pre-test and the post-test were analyzed for the 18 fake news items and compared between conditions. For reasons that are not entirely clear, the three true news items were not mentioned nor analyzed in their paper. The results showed that, compared to participants in the control condition, those in the treatment condition demonstrated a greater decrease in reliability ratings to fake news items from the pre-test to the post-test. Consequently, Basol et al. (2020) concluded that Bad News improved participants' ability to spot misinformation and that their data "demonstrated the efficacy of a 'broad-spectrum' inoculation against misinformation" (p. 5).

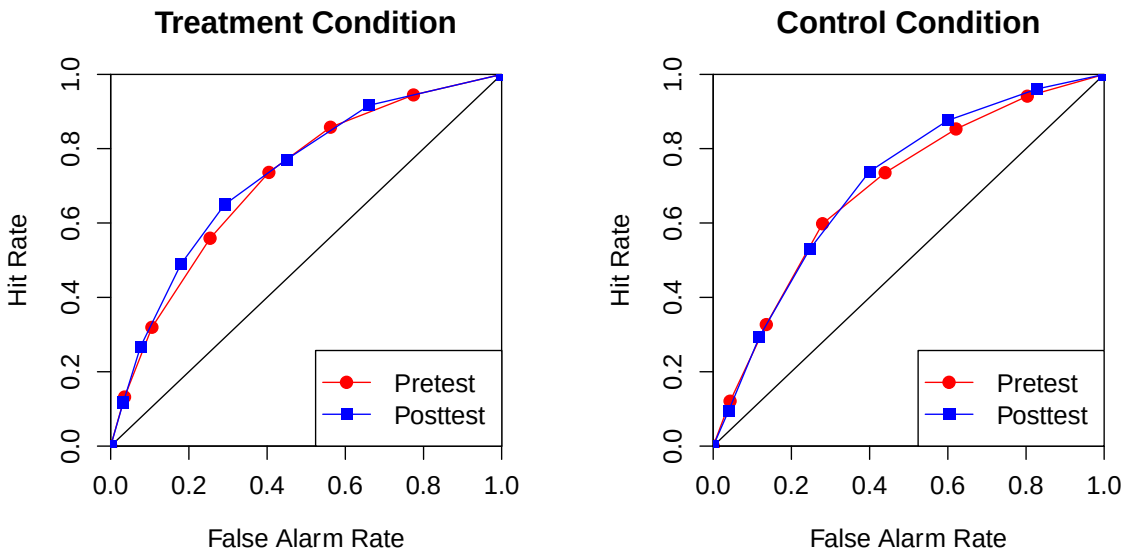
Basol et al. (2020) also collected data on how confident participants were in their reliability ratings. We did not reanalyze those data because our aim was to determine the effectiveness of gamified inoculation interventions for improving news veracity discernment, and reliability ratings on their own serve this purpose. Furthermore, although Basol et al. excluded the true news items from their analysis of mean ratings, it was necessary for us to include both true and fake news items in our reanalysis to assess discernment and response bias. Overall, Basol et al.'s full sample of 198 participants (102 from the control condition and 96 from the treatment condition) was used for our reanalysis.

Results

Treatment Condition. The ROC curves for the pre-test and the post-test in the treatment condition are shown in Figure 5. A paired samples *t*-test revealed that the AUC values for the pre-test (M = .73, SD = .18) were not significantly different from the AUC values for the post-test (M = .75, SD = .20), $t(95) = 0.92$, $p = .359$, $d = 0.09$, 95% CI [-0.11, 0.29], and the Bayes factor indicated moderate evidence for the null hypothesis, $BF_{10} = 0.17$. Both the HRs and the FARs decreased between the pre-test and the post-test (see Table S3). A paired samples *t*-test revealed that the *B"D* values for the pre-test (M = .07, SD = .78) were significantly smaller than the *B"D* values for the post-test (M = .24, SD = .78), $t(95) = 5.54$, $p < .001$, $d = 0.47$, 95% CI [0.29, 0.64] (see Table S4), and the Bayes factor indicated extreme evidence in favor of the alternative hypothesis, $BF_{10} = 48{,}915.96$.

Figure 5

ROC Curves for the Pre-Test and Post-Test in Basol et al.'s (2020) Treatment Condition and Control Condition



Note. ROC = Receiver operating characteristic.

Control Condition. The ROC curves for the pre-test and the post-test in the control condition are shown in Figure 5. A paired samples *t*-test revealed that the AUC values for the pre-test (M = .70, SD = .18) were not significantly different from the AUC values for the

post-test (M = .71, SD = .19), t(101) = 0.61, p = .543, d = 0.05, 95% CI [-0.11, 0.21], and the Bayes factor indicated moderate evidence for the null hypothesis, BF_{10} = 0.13. Both the HRs and the FARs only slightly decreased between the pre-test and the post-test (see Table S5). A paired samples t-test revealed that the $B''D$ values for the pre-test (M = .03, SD = .77) were not significantly different from the $B''D$ values for the post-test (M = .07, SD = .79), t(101) = 1.71, p = .090, d = 0.12, 95% CI [-0.02, 0.27] (see Table S6). The Bayes factor indicated anecdotal evidence for the null hypothesis, BF_{10} = 0.45.

Discussion

In summary, neither Bad News nor Tetris improved participants' news veracity discernment. These results stand in contrast to those of our previous reanalysis of Roozenbeek and van der Linden's (2019) data in which discernment was better after playing Bad News compared to before. We hypothesized that the improved discernment in the previous reanalysis may have been an artifact of using true news items that were obviously true. Indeed, several factors point to this factor being critical. Firstly, compared to Roozenbeek and van der Linden (2019), the shape of the ROC curve was notably less bowed in Basol et al. (2020), and the AUC values were considerably smaller (AUC in Roozenbeek and van der Linden's data: .88–.90; AUC in Basol et al.'s data: .70–.75). Moreover, this difference in AUC was primarily due to much lower HRs to true news items in Basol et al.'s data. Most notably, the HR associated with scale value 7 in Roozenbeek and van der Linden's data was .60, whereas it was only .09–.12 in Basol et al.'s study. In other words, participants assigned the most extreme reliability rating to the true news items 60% of the time in Roozenbeek and van der Linden's study, whereas this type of responding only occurred about 10% of the time in Basol et al.'s study – a sixfold difference. A closer look at the items used in Basol et al.'s study reveals why this occurred: Unlike the items used in Roozenbeek and van der Linden's study, the true and fake items were similarly obscure (e.g., "Super Bowl overnight TV ratings hit 10-year low" [true], "The 8th season of #GameOfThrones will be postponed due to a salary dispute" [fake]). Consistent with this observation, the overall average pre-test reliability rating collapsed across the two conditions

was 4.57 for the three true news items and 3.23 across the 18 fake news items, neither of which are close to the lower or upper limit of the scale (i.e., 1 or 7).

While this reanalysis showed that Bad News did not improve discernment, it *did* influence response bias. That is, the *B"D* analysis showed that after playing Bad News (but not Tetris), participants rated both true and fake news items as less reliable. Basol et al., by contrast, did not report any analyses of the true news items, and only reported that playing Bad News resulted in a greater decrease in reliability ratings to fake news items than playing Tetris. These findings are consistent with the response bias effect we report here. However, because true news items were omitted from their analyses, this difference in response bias was misinterpreted to be “clear evidence in support of the intervention” (Basol et al., 2020, p. 5), a conclusion that we do not believe is supported by their data.

Roozenbeek et al. (2020)

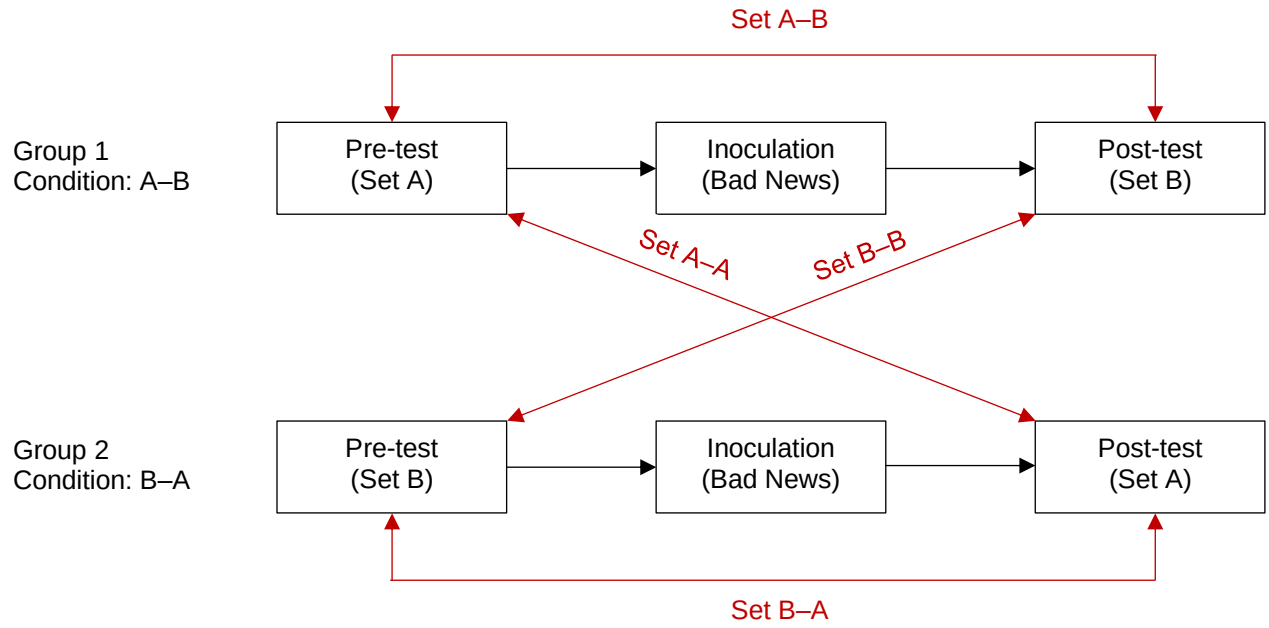
The goal of Roozenbeek et al.’s (2020) study was to address two methodological issues associated with using pre-post designs to assess the effectiveness of Bad News. The first issue pertains to item effects, which can result from presenting the same items in the pre-test and the post-test, as this raises the concern that any observed effects are specific to those particular items. The second issue pertains to testing effects, which can result from the implementation of a pre-test, as prior experience with the testing procedure could be a cause of any observed effects. Roozenbeek et al. conducted Experiment 1 to examine item effects and Experiment 2 to examine testing effects. For both experiments, a voluntary pre-post survey was embedded in Bad News that asked players who opted in to rate the reliability of news items on a scale that ranged from 1 (*not at all reliable*) to 7 (*very reliable*).

In Experiment 1, two different sets of news items labelled Set A and Set B were used. Both sets contained a total of eight news items that were presented in the form of Twitter posts. Six of these news items were created by the researchers and contained false information, and two were obtained from global news events and contained true information. There was one fake news item for each misinformation technique presented in Bad News. Participants were randomly assigned to one of two conditions: A–B, where participants were

presented with Set A in the pre-test and Set B in the post-test, and B–A, where participants were presented with Set B in the pre-test and Set A in the post-test. To test for item effects, the four different sets of pre-post differences shown in Figure 6 were analyzed.

Figure 6

Design of Roozenbeek et al.'s (2020) Experiment 1 and the Pre-Post Differences of Interest



Note. The pre-post differences of interest are indicated by the red arrows. Adapted from "Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation: Solomon Revisited", by J. Roozenbeek, R. Maertens, W. McClanahan, S. van der Linden, 2020, *Educational and Psychological Measurement*, *81*(2), pp. 340–362 (https://doi.org/10.1177%2F0013164420940378).

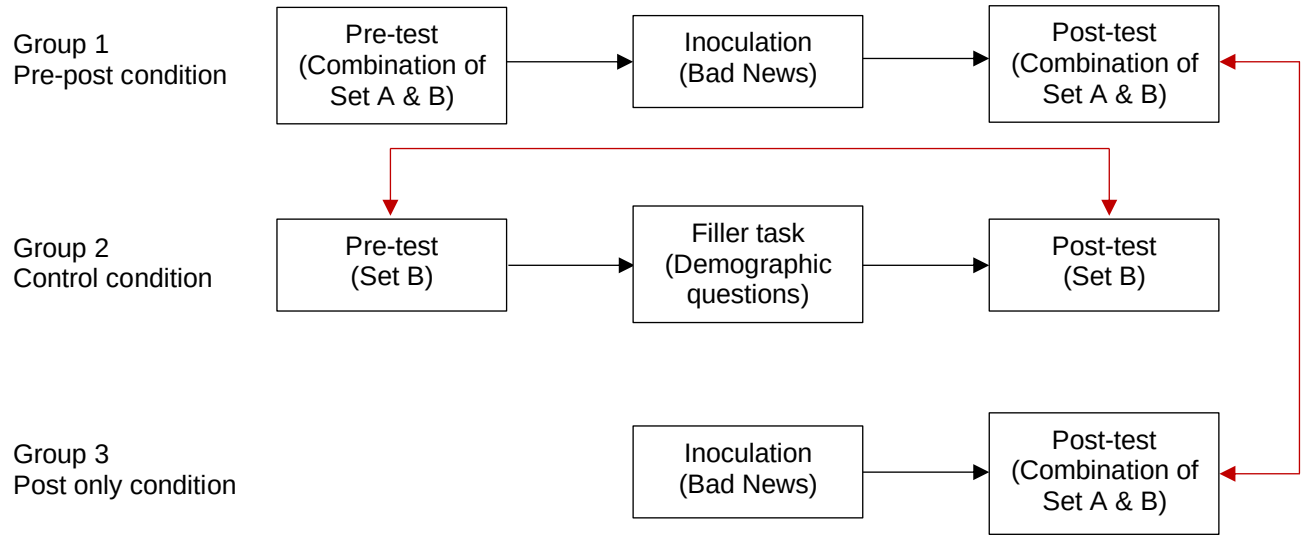
Roozenbeek et al. (2020) argued that item effects would be indicated through significant disparities in pre-post differences between either Set A–A and Set B–B, or Set A–B and Set B–A. The results showed that pre-post differences were significant for Set A–A but not Set B–B. Furthermore, although pre-post differences were significant for both Set A–B and Set B–A, after using standardized tests (i.e., the same tests but on z-scores based on the means and standard deviations of the pre-test scores for each item set) to account for

the different items being used in the pre-test and the post-test, the pre-post differences for Set A–B were no longer significant, whereas the pre-post differences for Set B–A remained significant. Overall, these findings suggested that there were indeed item effects.

In Experiment 2, participants were randomly assigned to one of three conditions: (a) pre-post, where participants first completed a pre-test, then played Bad News, and then completed a post-test; (b) control, where participants first completed a pre-test, then did a filler task (demographic questions), and then completed a post-test; and (c) post only, where participants first played Bad News and then completed a post-test. Although Set B was used for the control condition, a combination of Set A and Set B items was used for the post only and pre-post conditions. This combination still amounted to two real news items and six fake news items that reflected the six misinformation techniques presented in Bad News (one technique per item). Roozenbeek et al. (2020) argued that testing effects would be indicated by significant pre-post differences in the control condition, and a significant difference between post-test scores in the pre-post and post only conditions (see Figure 7). The results suggested that there were no testing effects since these differences were not significant.

Figure 7

Design of Roozenbeek et al.'s (2020) Experiment 2 and the Differences of Interest



Note. The differences of interest are indicated by the red arrows. Adapted from "Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation: Solomon Revisited", by J. Roozenbeek, R. Maertens, W. McClanahan, S. van der Linden, 2020, *Educational and Psychological Measurement*, *81*(2), pp. 340–362 (https://doi.org/10.1177%2F0013164420940378).

Due to the item effects demonstrated in Roozenbeek et al.'s (2020) study, we limited our reanalysis to pre-post differences between equivalent item sets. Therefore, we reanalyzed: (a) Set A–A in Experiment 1; (b) Set B–B in Experiment 1; and (c) the control condition in Experiment 2. Overall, data from 1240 participants (480 from both Set A–A and Set B–B, and 760 from the control condition) were used for our reanalysis.

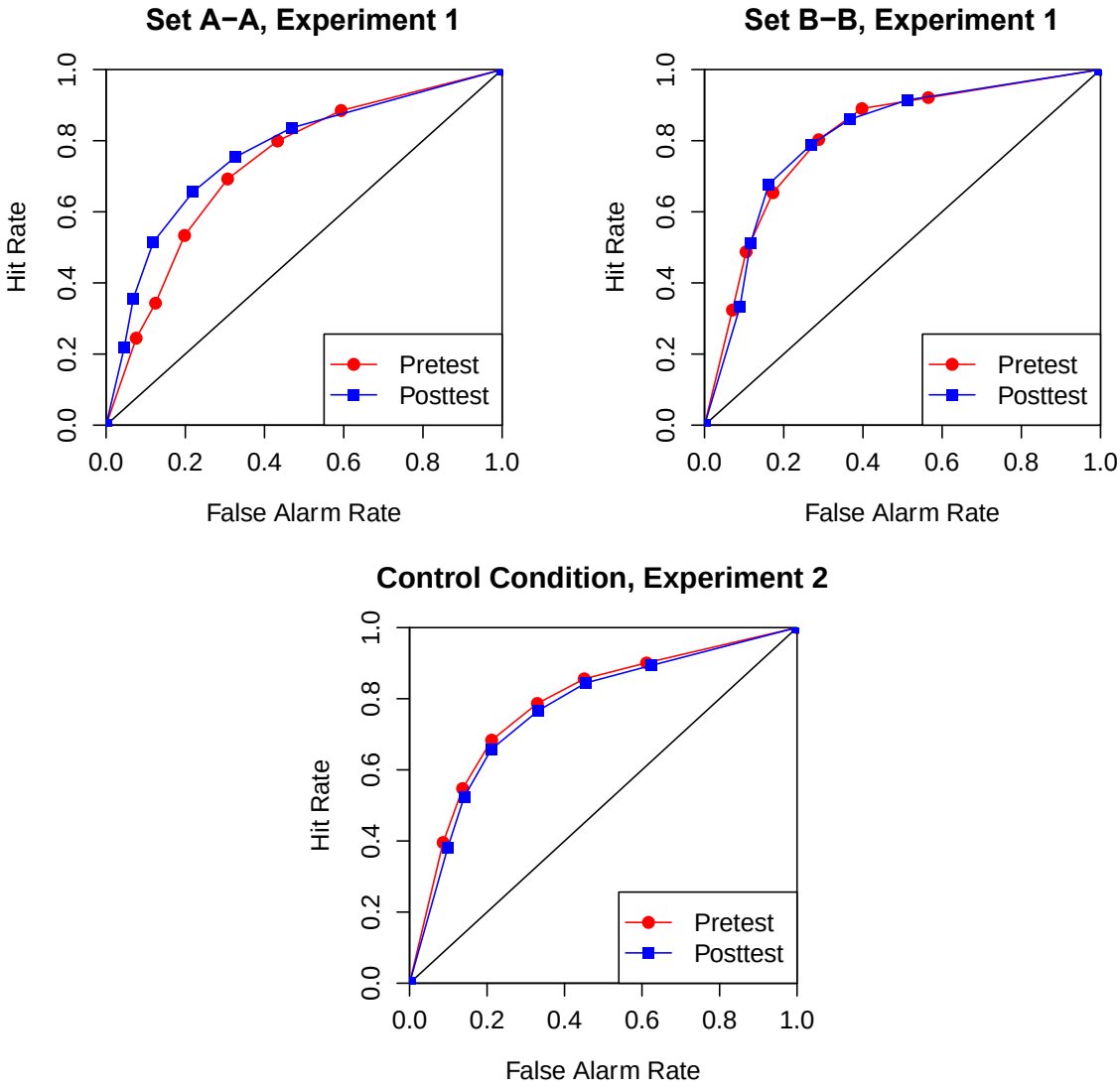
Results

Experiment 1: Set A–A. The ROC curves for the pre-test and the post-test from Set A–A in Experiment 1 are shown in Figure 8. Despite the apparent separation of the ROC curves, a Welch's independent samples *t*-test revealed that the AUC values for the pre-test (M = .75, SD = .22) were not significantly different from the AUC values for the post-test (M = .78, SD = .18), $t(462.62) = 1.79$, $p = .074$, $d = 0.16$, 95% CI [-0.02, 0.34]. The Bayes factor indicated anecdotal evidence for the null hypothesis, $BF_{10} = 0.48$. The FARs decreased between the pre-test and the post-test, and so did the HRs, albeit to a lesser degree (see

Table S7). A Welch's independent samples *t*-test revealed that the *B"D* values for the pre-test (M = .21, SD = .69) were significantly smaller than the *B"D* values for the post-test (M = .36, SD = .64), $t(476.32) = 4.32$, $p < .001$, $d = 0.39$, 95% CI [0.21, 0.58] (see Table S8), and the Bayes factor indicated extreme evidence in favor of the alternative hypothesis, BF_{10} = 764.51.

Figure 8

ROC Curves for the Pre-Test and Post-Test in Roozenbeek et al.'s (2020) Set A–A and Set B–B From Experiment 1 and Control Condition From Experiment 2



Note. ROC = Receiver operating characteristic.

Experiment 1: Set B–B. The ROC curves for the pre-test and the post-test from Set B–B in Experiment 1 are shown in Figure 8. A Welch's independent samples *t*-test revealed that the AUC values for the pre-test (M = .83, SD = .18) were not significantly different from the AUC values for the post-test (M = .84, SD = .19), $t(476.10) = 0.75$, p = .452, d = 0.07, 95% CI [-0.11, 0.25], and the Bayes factor indicated moderate evidence for the null hypothesis, BF_{10} = 0.13. The HRs remained mostly equal across the pre-test and the post-test, while the FARs only slightly decreased (see Table S9). A Welch's independent samples *t*-test revealed that the *B"D* values for the pre-test (M = .14, SD = .67) were not significantly different from the *B"D* values for the post-test (M = .17, SD = .69), $t(456.95) = 0.80$, p = .426, d = 0.07, 95% CI [-0.11, 0.25] (see Table S10). The Bayes factor indicated moderate evidence in favor of the null hypothesis, BF_{10} = 0.14.

Experiment 2: Control Condition. The ROC curves for the pre-test and the post-test from the control condition in Experiment 2 are shown in Figure 8. A paired samples *t*-test revealed that although the AUC values for the pre-test (M = .80, SD = .20) were significantly greater than the AUC values for the post-test (M = .79, SD = .21), the effect size was almost negligible, $t(759) = -2.09$, p = .037, d = -0.07, 95% CI [-0.13, -0.00]. The Bayes factor indicated anecdotal evidence for the null hypothesis, BF_{10} = 0.36. The HRs only slightly decreased across the pre-test and the post-test, while the FARs only slightly increased (see Table S11). A paired samples *t*-test revealed that the *B"D* values for the pre-test (M = .06, SD = .68) were not significantly different from the *B"D* values for the post-test (M = .08, SD = .71), $t(759) = 1.68$, p = .094, d = 0.05, 95% CI [-0.01, 0.11] (see Table S12), and the Bayes factor indicated moderate evidence for the null hypothesis, BF_{10} = 0.16.

Discussion

Overall, participants in the control condition in Experiment 2, where the Bad News intervention was absent, did not demonstrate a change in news veracity discernment or response bias from the pre-test to the post-test. When Bad News intervened between the pre-test and post-test in Experiment 1, it did not improve participants' news veracity discernment in either Set A–A or Set B–B, but it did elicit a more conservative response bias

in Set A–A. This is noteworthy since Set B–B and the control condition used the same set of news items to assess participants' reliability ratings (i.e., Set B), whereas Set A–A used a different set of news items (i.e., Set A). These results could potentially be attributed to differences in ambiguity between the true versus fake news items used in each set, as those in Set A (AUC = .75–.78) were more ambiguous than those in Set B (AUC = .79–.84). Indeed, the mean reliability ratings support this conclusion. Specifically, the mean pre-test reliability rating was 4.50 for the true news items and 2.73 for the fake news items in Set A–A (difference = 1.77), whereas the respective mean pre-test ratings were 5.08 and 2.60 for true and fake news items in Set B–B (difference = 2.48).

Thus, analogous to the reanalysis of Basol et al.'s (2020) data, when ambiguous news items were used (i.e., Set A), Bad News prompted more conservative responding on the post-test compared to the pre-test. Conversely, when the items were less ambiguous such that participants had strong opinions about their objective veracity on the pre-test (i.e., Set B), Bad News had little effect on response bias. More specifically, the *B"D* analysis showed that playing Bad News caused participants to become more conservative in their responses to news items in Set A–A, but not in Set B–B. Although both the HRs (true news) and FARs (fake news) were lower in the post-test than the pre-test for Set A, the decrease was greater for the FARs. This finding explains the significant pre-post differences in reliability ratings for the fake news items but not the true news items in Set A–A reported by Roozenbeek et al. (2020). Furthermore, the fact that both playing Bad News and doing nothing (control) had null effects on both news veracity discernment and response bias for Set B items is consistent with the non-significant pre-post differences in reliability ratings in Set B–B reported by Roozenbeek et al. Ultimately, our findings suggest that Bad News can cause participants to respond more conservatively but does not improve discernment, and that this shift to more conservative responding is more likely to occur when the news items used to assess reliability ratings are more ambiguous.

Maertens et al. (2021)

Maertens et al. (2021) conducted three experiments to investigate the long-term effectiveness of Bad News. In Experiments 1 and 2, participants were randomly allocated to either a control condition that played Tetris or a treatment condition that played Bad News. In Experiment 3, all participants played Bad News. All three experiments included a pre-test and several post-tests that required participants to rate the reliability of news items on a scale ranging from 1 (*not at all reliable*) to 7 (*very reliable*). In Experiment 1, the pre-test occurred just before playing either Tetris or Bad News, and the post-tests were administered immediately, 1 week, 5 weeks, and 13 weeks after the pre-test. In Experiments 2 and 3, the pre-test and the first post-test followed the same scheduling as in Experiment 1, but there was only one follow-up post-test. This follow-up post-test occurred 9 weeks after the pre-test in Experiment 2 and 1 week after the pre-test in Experiment 3.

In Experiments 1 and 2, the news items did not vary between the pre-test and the post-tests. In both experiments, participants were presented with 21 news items in each test, 18 of which were created by the researchers and contained false information, and three of which were obtained from global news events and contained true information. The same 18 fake news items were used in Experiments 1 and 2, but two out of the three true news items were different. There were three fake news items corresponding to each of the six misinformation techniques presented in Bad News. In Experiment 3, the ratio of true to fake news items was changed to 1:6 in the pre-test and the first post-test, and to 6:6 in the second post-test. The news items in the pre-test and the first post-test differed from the news items in the second post-test. Nevertheless, every pre-test and post-test had one fake news item for every misinformation technique. Furthermore, the fake news items were created by the researchers while the true news items were obtained from global news events in the same way as in Experiments 1 and 2. In all three experiments, the news items were presented in the form of news headlines. Critically, the true news items were excluded from the main analysis for all three experiments.

In Experiment 1, participants in the treatment condition demonstrated a significantly greater decrease in mean fake news reliability ratings from the pre-test to all four post-tests

than participants in the control condition. In Experiment 2, participants in the treatment condition demonstrated a significantly greater decrease in fake news reliability ratings from the pre-test to the first post-test, but not the second post-test, than participants in the control condition. In Experiment 3, participants demonstrated a significant decrease in fake news reliability ratings from the pre-test to the two post-tests. The results of Experiment 1 indicated that Bad News improved people's ability to identify misinformation for up to 13 weeks after playing the game. However, the results of Experiment 2 revealed that this was most likely due to repeated testing, and that the effect of Bad News decays and becomes negligible after 9 weeks. Finally, the results of Experiment 3 ruled out the possibility that the findings from Experiments 1 and 2 were due to the unbalanced ratio of true to fake news headlines and/or the presentation of the same news headlines in the pre-test and the post-tests.

In Experiment 3, only one true news item was used in the pre-test and the first post-test compared to six in the follow-up post-test. In our view, a single true news item in the pre-test and the first post-test would not allow for a representative account of participants' belief in true news. As a result, we limited our reanalysis to Experiments 1 and 2. Furthermore, although Maertens et al. (2021) excluded the true news items from their main analysis and only included the fake news items, it was necessary for us to include both types of news items in our reanalysis so that ROC curves could be constructed. Finally, we limited our reanalysis to participants who had completed the entire experiment. This resulted in a sample of 118 participants for Experiment 1 (58 from the treatment condition and 60 from the control condition), and 110 for Experiment 2 (54 from the treatment condition and 56 from the control condition).

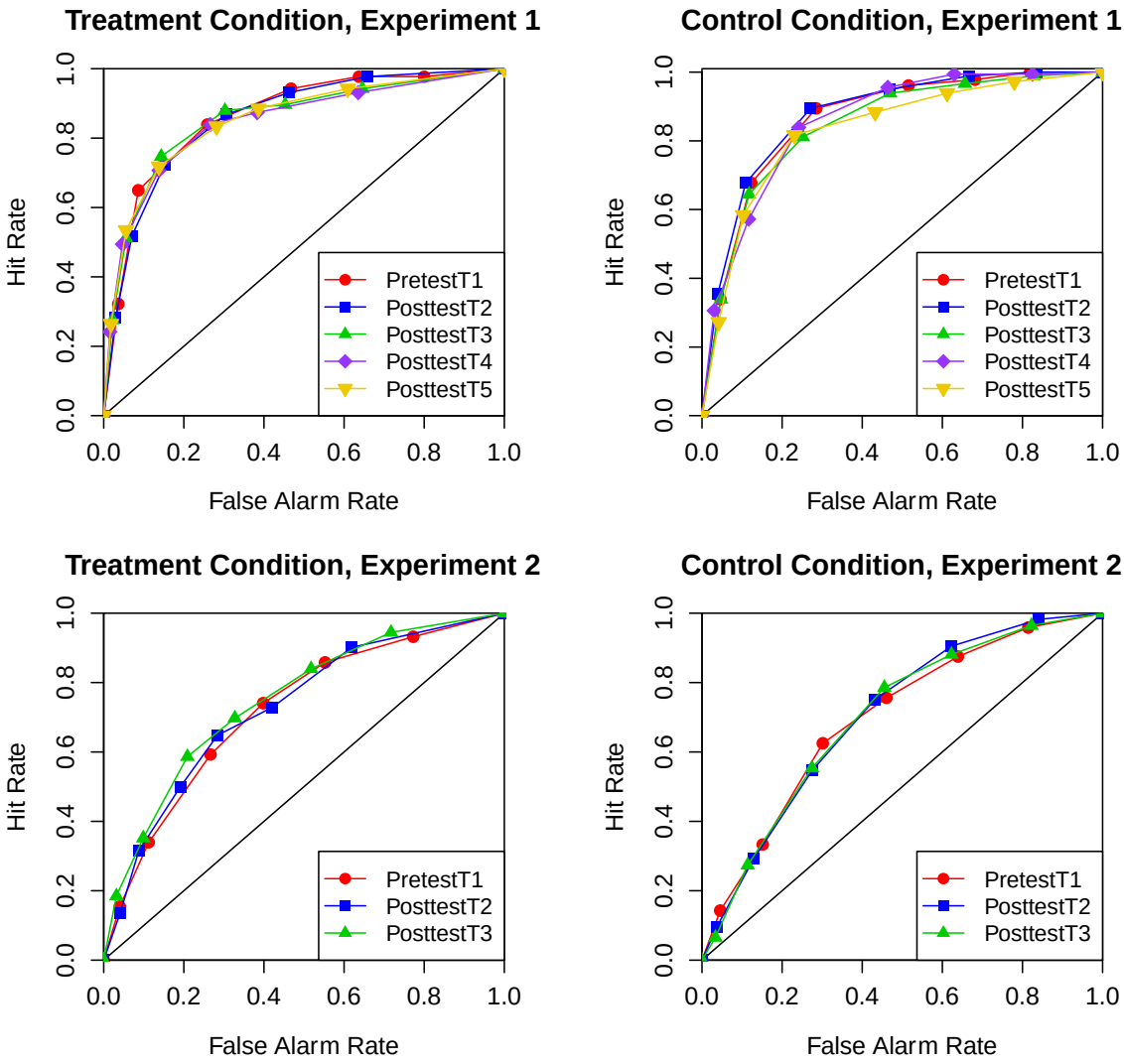
Results

Experiment 1: Treatment Condition. The ROC curves for the pre-test and the four post-tests from the treatment condition in Experiment 1 are shown in Figure 9. A one-way repeated-measures ANOVA revealed that the main effect of test on AUC values was not significant, $F(4, 228) = 0.19$, $p = .943$, $\eta^2 = .00$, 95% CI [.00, .01], and the Bayes factor

indicated very strong evidence for the null hypothesis, BF_{10} = 0.01. The means and standard deviations of the AUC values are shown in Table S13. Both the HRs and the FARs decreased between the pre-test and the four post-tests but varied only slightly between the four post-tests (see Table S14). A one-way repeated-measures ANOVA revealed that the main effect of test on *B"D* values was significant, $F(4, 228) = 27.70$, $p < .001$, $\eta^2 = .33$, 95% CI [.23, .41], and the Bayes factor indicated extreme evidence in favor of the alternative hypothesis, $BF_{10} = 5.32 \times 10^{15}$ (see Table S15). A post-hoc Tukey test showed that the pre-test was significantly different from each of the four post-tests at $p < .001$, but none of the four post-tests were significantly different from each other, smallest $p = .059$.

Figure 9

ROC Curves for the Pre-Test and Post-Tests in Maertens et al.'s (2021) Treatment Condition and Control Condition From Experiment 1 and Treatment Condition and Control Condition From Experiment 2



Note. ROC = Receiver operating characteristic. T = time. T1 = just before playing Bad News or Tetris in Experiments 1 and 2. T2 = just after playing Bad News or Tetris in Experiments 1 and 2. T3 = 1 week after playing Bad News or Tetris in Experiment 1 and 9 weeks after playing Bad News or Tetris in Experiment 2. T4 = 5 weeks after playing Bad News or Tetris in Experiment 1. T5 = 13 weeks after playing Bad News or Tetris in Experiment 1.

Experiment 1: Control Condition. The ROC curves for the pre-test and the four post-tests from the control condition in Experiment 1 are shown in Figure 9. A one-way

repeated measures ANOVA revealed that the main effect of test on AUC values was not significant, $F(4, 236) = 0.79$, $p = .534$, $\eta^2 = .01$, 95% CI [.00, .04], and the Bayes factor indicated very strong evidence for the null hypothesis, $BF_{10} = 0.03$. The means and standard deviations of the AUC values are shown in Table S16. Both the HRs and the FARs decreased between the pre-test and the final post-test but varied only slightly between the pre-test and the first three post-tests (see Table S17). A one-way repeated measures ANOVA revealed that the main effect of test on *B"D* values was significant, $F(4, 236) = 4.25$, $p = .002$, $\eta^2 = .07$, 95% CI [.01, .13], and the Bayes factor indicated moderate evidence in favor of the alternative hypothesis, $BF_{10} = 7.81$ (see Table S18). A post-hoc Tukey test showed that the pre-test was significantly different from the final post-test at $p = .037$, but none of the other comparisons was significant, smallest $p = .080$.

Experiment 2: Treatment Condition. The ROC curves for the pre-test and the two post-tests from the treatment condition in Experiment 2 are shown in Figure 9. A one-way repeated measures ANOVA revealed that the main effect of test on AUC values was not significant, $F(2, 106) = 0.74$, $p = .482$, $\eta^2 = .01$, 95% CI [.00, .07], and the Bayes factor indicated moderate evidence for the null hypothesis, $BF_{10} = 0.12$. The means and standard deviations of the AUC values are shown in Table S19. Both the HRs and the FARs decreased between the pre-test and the first post-test. The FARs also decreased between the pre-test and the second post-test, but to a lesser degree than with the first post-test, whereas the HRs remained static (see Table S20). A one-way repeated measures ANOVA revealed that the main effect of test on *B"D* values was significant, $F(2, 106) = 10.45$, $p < .001$, $\eta^2 = .16$, 95% CI [.05, .29], and the Bayes factor indicated extreme evidence in favor of the alternative hypothesis, $BF_{10} = 289.85$ (see Table S21). A post-hoc Tukey test showed that the pre-test was significantly different from the first post-test at $p < .001$, but none of the other comparisons was significant, smallest $p = .060$.

Experiment 2: Control Condition. The ROC curves for the pre-test and the two post-tests from the control condition in Experiment 2 are shown in Figure 9. A one-way repeated measures ANOVA revealed that the main effect of test on AUC values was not

significant, $F(2, 110) = 0.10$, $p = .909$, $\eta^2 = .00$, 95% CI [.00, .02], and the Bayes factor indicated very strong evidence for the null hypothesis, $BF_{10} = 0.06$. The means and standard deviations of the AUC values are shown in Table S22. Both the HRs and the FARs only slightly decreased between the pre-test and the two post-tests (see Table S23). A one-way repeated measures ANOVA revealed that the main effect of test on *B"D* values was not significant, $F(2, 110) = 0.62$, $p = .541$, $\eta^2 = .01$, 95% CI [.00, .07] (see Table S24), and the Bayes factor indicated moderate evidence in favor of the null hypothesis, $BF_{10} = 0.10$.

Discussion

In summary, neither Bad News nor Tetris improved participants' news veracity discernment in either Experiment 1 or 2. In Experiment 1, Bad News elicited more conservative responding immediately after playing, as well as 1, 5, and 13 weeks later, and Tetris elicited more conservative responding 13 weeks after playing. In Experiment 2, Bad News elicited more conservative responding immediately after playing but not 9 weeks later, whereas Tetris did not elicit more conservative responding at all. Although the results from Experiment 1 showed that both Bad News and Tetris caused participants to respond more conservatively, they were confounded by the effects of repeated testing. However, this confound had equal influence on the treatment condition and the control condition, and Bad News caused an increase in conservative responding at each post-test, while Tetris only caused an increase in conservative responding at one post-test. Therefore, even if both games caused more conservative responding, the one produced by Bad News was more consistent in Experiment 1. Moreover, Experiment 2 was not confounded by the effects of repeated testing, and it showed that Bad News increased conservative responding, whereas Tetris did not.

Although Maertens et al. (2021) excluded the true news items from the main analysis, our reanalysis allows us to add nuance to their results. In Experiment 1, Bad News increased conservative responding at all four post-tests, while Tetris only did so at the final post-test. This explains the significantly greater decrease in reliability ratings from the pre-test to all four post-tests in the treatment condition compared to the control condition. In

Experiment 2, Bad News increased conservative responding at the first post-test but not the second, while Tetris did not do so at all. This explains the significantly greater decrease in mean reliability ratings from the pre-test to the first post-test in the treatment condition compared to the control condition, as well as the non-significant difference in reliability ratings from the pre-test to the second post-test in both conditions. Ultimately, our findings suggest that Bad News can cause more conservative responding but not an improvement in discernment, and that this change to response bias disappears over time unless participants undergo repeated testing.

Reanalysis of Go Viral!

Basol et al. (2021)

The aim of Basol et al.'s (2021) paper was to investigate the effectiveness of Go Viral! and a series of infographics on people's ability to identify COVID-19 misinformation. For this purpose, they conducted two different studies. In Study 1, a voluntary pre-post survey was embedded in Go Viral! that asked players who opted in to rate the manipulativeness of the same six news items before and after playing the game. Manipulativeness ratings were made on a scale that ranged from 1 (*not at all manipulative*) to 7 (*very manipulative*), and the six news items were presented in the form of Twitter posts. Half of these news items were obtained from fact-checking websites and contained false information, and the other half were obtained from the Twitter accounts of reputable news sources and contained true information. The three fake news items reflected the three misinformation techniques presented in Go Viral! (one technique per item). Source information was omitted from each news item to prevent participants from only using this feature to spot misinformation.

To test whether Go Viral! improved participants' ability to spot misinformation, differences in mean manipulativeness ratings between the pre-test and the post-test were analyzed for the six news items. The results showed that participants rated the fake news items as significantly more manipulative in the post-test compared to the pre-test, but there were no significant pre-post differences for the true news items. To test whether Go Viral!

improved participants' news veracity discernment, differences in mean manipulativeness ratings between the true and fake news items were compared in the pre-test and the post-test. The results showed that this difference was significantly larger in the post-test compared to the pre-test. Overall, Basol et al. (2021) concluded that Go Viral! improved people's ability to spot misinformation as well as their ability to distinguish between true and fake news.

In Study 2, participants were randomly assigned to one of three conditions: (a) control, where participants played Tetris; (b) passive, where participants were shown a series of infographics; and (c) active, where participants played Go Viral!. All conditions included a pre-test and a post-test that asked participants to rate the manipulativeness of the same 18 news items before and after experiencing either Tetris, Go Viral!, or infographics. Manipulativeness ratings were made on the same scale as in Study 1. Furthermore, the 18 news items were collected and presented in the same way, had the same 1:1 ratio of true to fake news items, and incorporated the same misinformation techniques as in Study 1. Study 2 was conducted in English, German, and French, and participants who completed the study in English were asked to take part in a follow-up test 1 week after the initial test date. This follow-up test required participants to rate the manipulativeness of 12 different news items, half of which were true and half of which were fake.

To test whether Go Viral! or the infographics improved participants' ability to spot misinformation, differences in mean manipulativeness ratings between the pre-test and the post-test were analyzed for the 18 news items and compared between conditions. The results showed that participants in the active and passive conditions rated the fake news items as more manipulative than participants in the control condition. However, participants in the active condition also rated the true news items as more manipulative than participants in the passive and control conditions. To test whether the effects of Go Viral! or the infographics decayed over time, the mean manipulativeness ratings for the 12 news items in the 1-week follow-up test were compared between conditions. The results showed that participants in the active condition rated the fake news items as more manipulative than

participants in the passive and control conditions, but there were no significant differences between conditions for the true news items.

Overall, Basol et al. (2021) concluded that Go Viral! and the infographics improved participants' ability to spot misinformation, and that this effect lasted a week for Go Viral!, but not for the infographics. Critically, Go Viral! worsened participants' ability to spot true information, but this effect disappeared after a week.

Basol et al. also collected data on how confident participants were in their manipulativeness ratings as well as how willing they were to share the news items online. Here, we focused on the manipulativeness ratings as these are suitable for ROC analysis. Furthermore, although we reanalyzed both Study 1 and Study 2, we limited our reanalysis to the active and control conditions from Study 2 as the effectiveness of infographics was not our primary interest. Finally, we limited our reanalysis to participants who had completed the entire experiment. To keep the reanalyses consistent, we reverse-coded the manipulativeness scores so that 1 = "very manipulative" and 7 = "not at all manipulative". Thus, the direction of the manipulativeness scale matched that of the reliability scales used in previous studies (i.e., higher values are associated with greater perceived truth). Overall, we used a sample of 1,771 participants for Study 1 and 386 participants for Study 2 (151 from the active condition and 235 from the control condition).

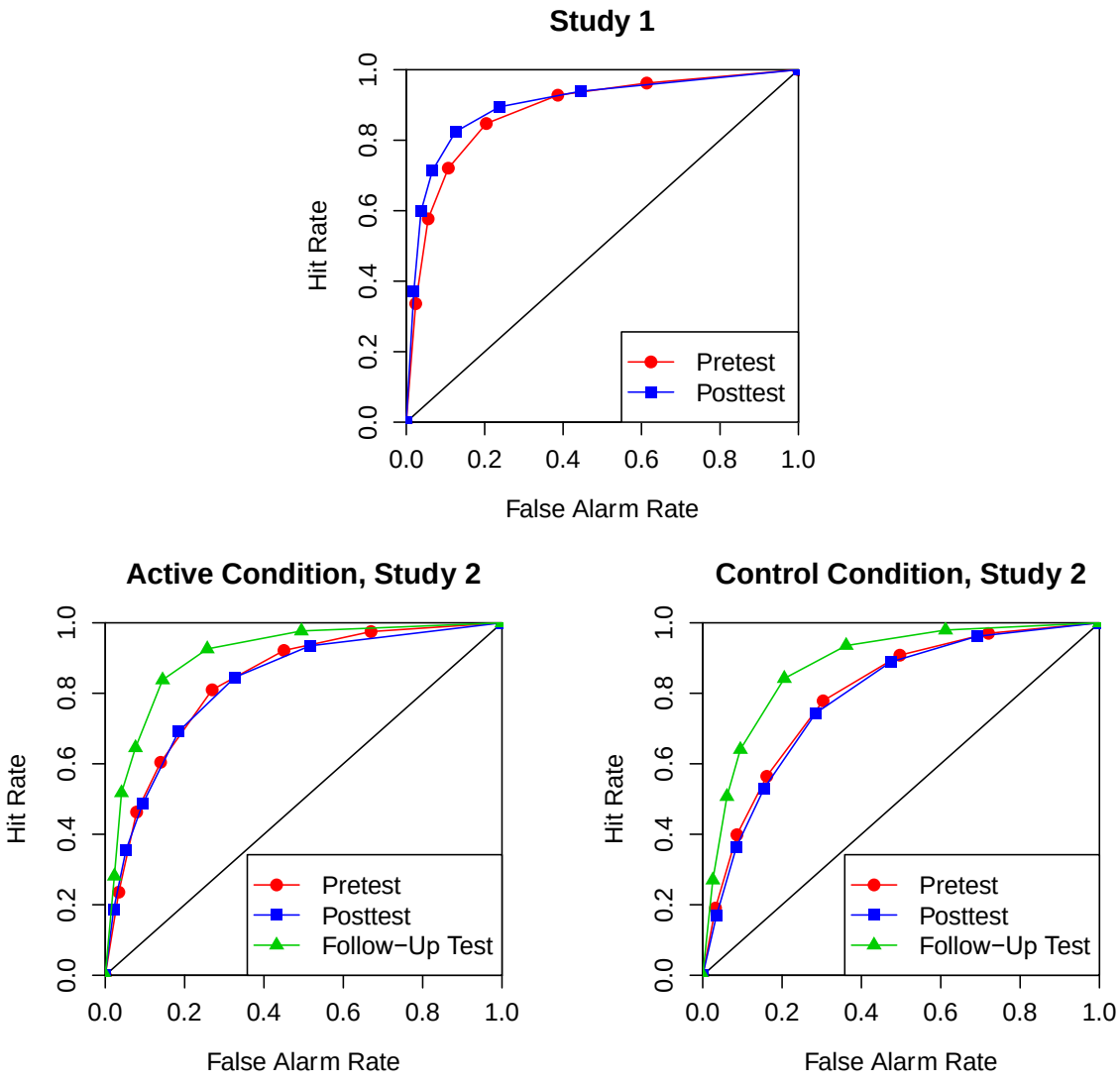
Results

Study 1. The ROC curves for the pre-test and the post-test in Study 1 are shown in Figure 10. A paired samples *t*-test revealed that although the AUC values for the pre-test (M = .89, SD = .18) were significantly smaller than the AUC values for the post-test (M = .90, SD = .18), the effect size was almost negligible, $t(1770) = 3.02$, $p = .003$, $d = 0.06$, 95% CI [0.02, .09], and the Bayes factor only indicated anecdotal evidence for the alternative hypothesis, $BF_{10} = 2.48$. The FARs decreased between the pre-test and the post-test, and so did the HRs, albeit to a lesser degree (see Table S25). A paired samples *t*-test revealed that the *B"D* values for the pre-test (M = .03, SD = .66) were significantly smaller than the *B"D* values for the post-test (M = .13, SD = .61), $t(1770) = 13.77$, $p < .001$, $d = 0.27$, 95% CI

[0.23, 0.31] (see Table S26), and the Bayes factor indicated extreme evidence in favor of the alternative hypothesis, $BF_{10} = 3.07 \times 10^{37}$.

Figure 10

ROC Curves for the Pre-Test and Post-Test in Basol et al.'s (2021) Study 1 and Active Condition and Control Condition From Study 2



Note. ROC = Receiver operating characteristic. The follow-up test took place 1 week after playing Go Viral! or Tetris in Study 2.

Study 2: Active Condition. The ROC curves for the pre-test, post-test, and follow-up test from the active condition in Study 2 are shown in Figure 10. A one-way repeated

measures ANOVA revealed that the main effect of test on AUC was significant, $F(2, 300) = 13.50$, $p < .001$, $\eta^2 = .08$, 95 % CI [.03, .14], and the Bayes factor indicated extreme evidence for the alternative hypothesis, $BF_{10} = 5045.25$. A post-hoc Tukey test showed that both the pre-test and the post-test were significantly different from the follow-up test at $p < .001$, but the pre-test was not significantly different from the post-test, $p = .973$. The means and standard deviations of the AUC values are shown in Table S27. Both the HRs and the FARs decreased between the pre-test and the post-test, whereas the HRs remained the same while the FARs decreased between the pre-test and the follow-up test (see Table S28). A one-way repeated measures ANOVA revealed that the main effect of test on *B"D* values was significant, $F(2, 300) = 26.01$, $p < .001$, $\eta^2 = .15$, 95% CI [.08, .22], and the Bayes factor indicated extreme evidence for the alternative hypothesis, $BF_{10} = 212{,}975{,}465$ (see Table S29). A post-hoc Tukey test showed that the pre-test was significantly different from both the post-test and the follow-up test at $p < .001$, and the post-test was significantly different from the follow-up test, $p = .004$.

Study 2: Control Condition. The ROC curves for the pre-test, post-test, and follow-up test from the control condition in Study 2 are shown in Figure 10. A one-way repeated measures ANOVA revealed that the main effect of test on AUC values was significant, $F(2, 468) = 11.60$, $p < .001$, $\eta^2 = .05$, 95 % CI [.02, .09], and the Bayes factor indicated extreme evidence for the alternative hypothesis, $BF_{10} = 860.03$. A post-hoc Tukey test showed that both the pre-test and the post-test were significantly different from the follow-up test, $p = .003$ and $p < .001$, respectively, but the pre-test was not significantly different from the post-test, $p = .361$. The means and standard deviations of the AUC values are shown in Table S30. Both the HRs and the FARs only slightly decreased between the pre-test and the post-test, whereas the HRs increased while the FARs decreased between the pre-test and the follow-up test (see Table S31). A one-way repeated measures ANOVA revealed that the main effect of test on *B"D* values was not significant, $F(2, 468) = 2.21$, $p = .111$, $\eta^2 = .00$, 95% CI [.00, .03], and the Bayes factor indicated moderate evidence for the null hypothesis, $BF_{10} = 0.13$ (see Table S32).

Discussion

Overall, in Study 1, Go Viral! elicited more conservative responding in the post-test compared to the pre-test, but its effect on participants' news veracity discernment was ambiguous. In Study 2, Go Viral! elicited more conservative responding but had no effect on news veracity discernment in the post-test compared to the pre-test. Although Tetris also did not have an effect on news veracity discernment, it did not elicit more conservative responding in the post-test compared to the pre-test. Finally, both Go Viral! and Tetris improved news veracity discernment in the follow-up test compared to the pre-test and the post-test. Specifically, Tetris decreased the FAR and increased the HR (cf. Tables S25 and S26), whereas Go Viral! decreased the FAR but barely affected the HR, which resulted in higher *B"D* values because the overall proportion of responses exceeding the upper criteria (i.e., those associated with scale values greater than 1) decreased (cf. Tables S23 and S24).

Considering that the delayed improvement in news veracity discernment was found in both the active condition and the control condition, coupled with the fact that different items were used in the follow-up test than in the pre-test and the post-test, this effect can be attributed to item differences. Specifically, the items used in the follow-up test appear to have been easier to discern than the items used in the pre-test and the post-test. This is supported by the average reliability ratings for the true and fake news items, respectively, collapsed across the two conditions; these were 4.97 and 2.60 in the pre-test, and 5.18 and 2.20 in the follow-up test, the latter of which are considerably closer to the lower and upper limits of the scale (i.e., 1 and 7).

These findings can be used to explain Basol et al.'s (2021) results. The *B"D* analysis showed that playing Go Viral! caused participants to respond more conservatively in Study 1. Although both the HRs (true news) and FARs (fake news) were lower in the post-test than the pre-test after playing Go Viral!, the decrease was greater for the FARs. This explains the significant pre-post differences in manipulativeness ratings for the fake news items but not the true news items in Study 1 of Basol et al. Similarly, Go Viral! caused participants to respond more conservatively in Study 2, while Tetris did not. This explains the larger pre-

post differences in mean manipulativeness ratings for both the true and the fake news items in the active condition compared to the control condition shown in Basol et al.'s paper. Finally, the HRs for the follow-up test were almost identical between the active condition and the control condition, whereas the FARs were slightly lower in the active condition compared to the control condition. This explains Basol et al.'s conclusion that participants who played Go Viral! rated fake news items as more manipulative than participants who played Tetris in the follow-up test.

Ultimately, our findings suggest that Go Viral! causes a shift to more conservative responding that can impact responses to both true and fake news items immediately after playing. However, its impact on news veracity discernment is unclear; it was ambiguous in Study 1, whereas in Study 2, it improved after 1 week, but this also occurred for participants that played Tetris. Therefore, considering Basol et al. (2021) is the only published paper that has examined the effectiveness of Go Viral!, further research on this topic is necessary to establish the relationship between playing Go Viral! and news veracity discernment.

Meta-Analysis

To examine the overall evidence for intervention-based effects on discernment and response bias across the five reanalyzed studies with k = 13 experiments, we conducted a meta-analysis on their pre-post AUC and *B"D* effect sizes, separately. For each set of pre-post effect sizes, we first fit two random-effects models, one on the k = 8 experiments that involved a treatment (i.e., Bad News or Go Viral!) and one on the k = 5 experiments that involved a control (i.e., Tetris or nothing). These two models compared the overall meta-analytic effect size estimate to the null effect size (i.e., d = 0). We then fitted a fixed-effects meta-regression model with a binary moderator variable (treatment/control) on the results from the two random-effects models.[6] This model compared the overall meta-analytic effect

[6] According to Borenstein et al. (2010), fixed-effects models assume that a common effect size underlies all the studies in the analysis, and that any observed effect size differences between them are due to sampling error. In contrast, random-effects models assume that a different effect size can underly each study due to heterogeneity between them. Considering the heterogeneity between the studies we analyzed (e.g., different participant pools, items, and treatment versions), we fitted random-effects models on the pre-post effect sizes. We then compared these two random-effects models with a fixed-effects model since the heterogeneity between studies had already been accounted for by the two prior random-effects models.

size estimate for the treatment conditions to the overall meta-analytic effect size estimate for the control conditions. This analysis was informed by the metafor R package (Viechtbauer, 2010) website (https://www.metafor-project.org/doku.php/tips:comp_two_independent_estimates).

The results of our meta-analysis on the pre-post AUC effect sizes are shown in Table 3 and Figure 11. The random-effects model on the treatment conditions showed a significant increase in AUC values from the pre-test to the post-test, while the random-effects model on the control conditions did not. However, the fixed-effects model comparing the two random-effects models showed no significant difference in pre-post AUC effect sizes between the treatment conditions and the control conditions. In other words, there was no overall evidence for intervention-based effects on discernment once the control data were taken into consideration.

Table 3

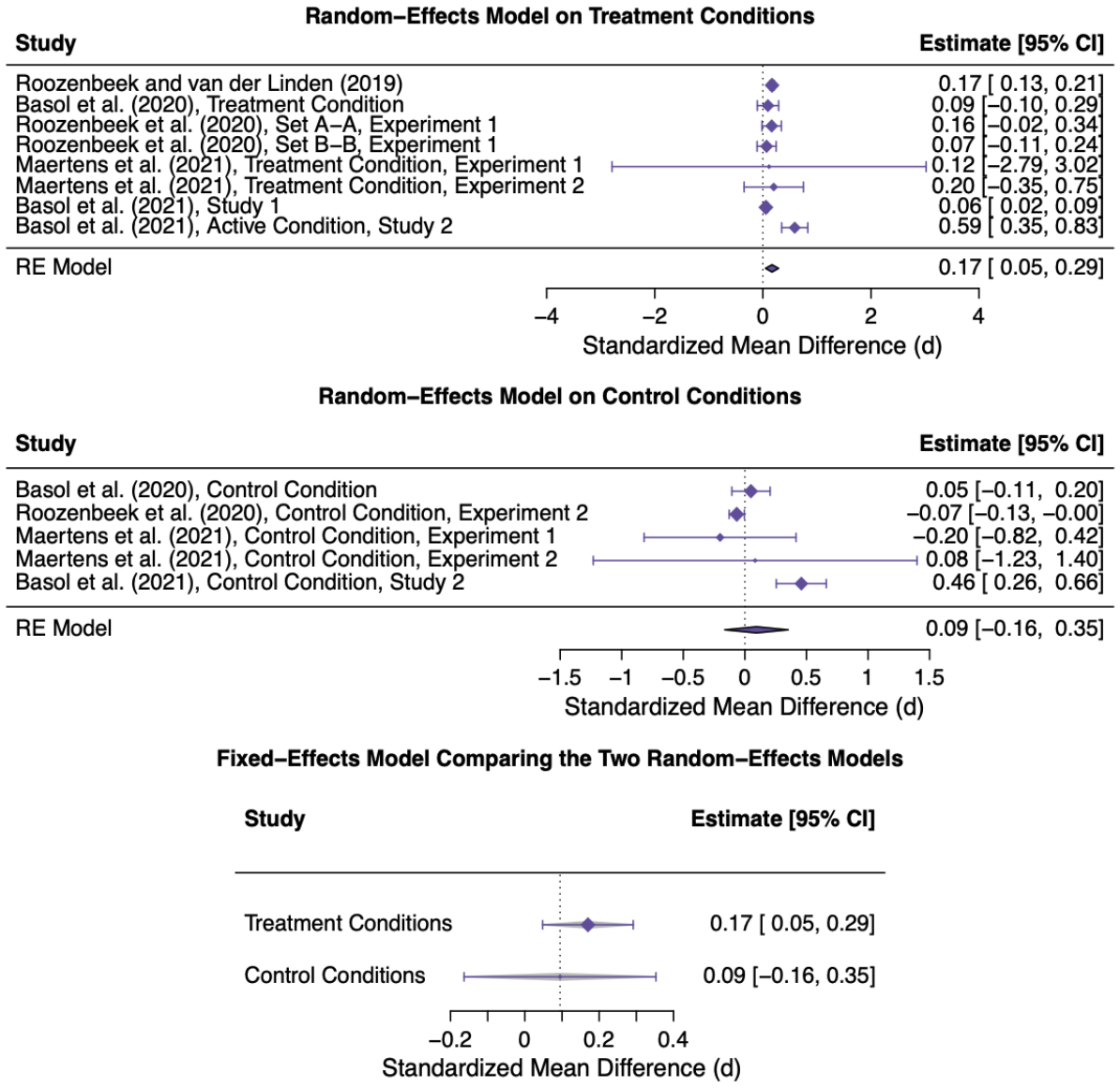
Results from the Meta-Analyses on the AUC and B"D Pre-Post Effect Sizes

Meta-analytic model	*d*	*z*	*p*
AUC pre-post effect sizes			
Random-effects model (treatment conditions)	0.17	2.73	.006
Random-effects model (control conditions)	0.09	0.72	.473
Fixed-effects model (comparing the two random-effects models)	0.08	0.52	.606
B"D pre-post effect sizes			
Random-effects model (treatment conditions)	0.57	3.92	< .001
Random-effects model (control conditions)	0.16	2.24	.025
Fixed-effects model (comparing the two random-effects models)	0.41	2.51	.012

Note. AUC = area under the curve.

Figure 11

The Two Random-Effects Models and the Fixed-Effects Model Fitted on the AUC Pre-Post Effect Sizes



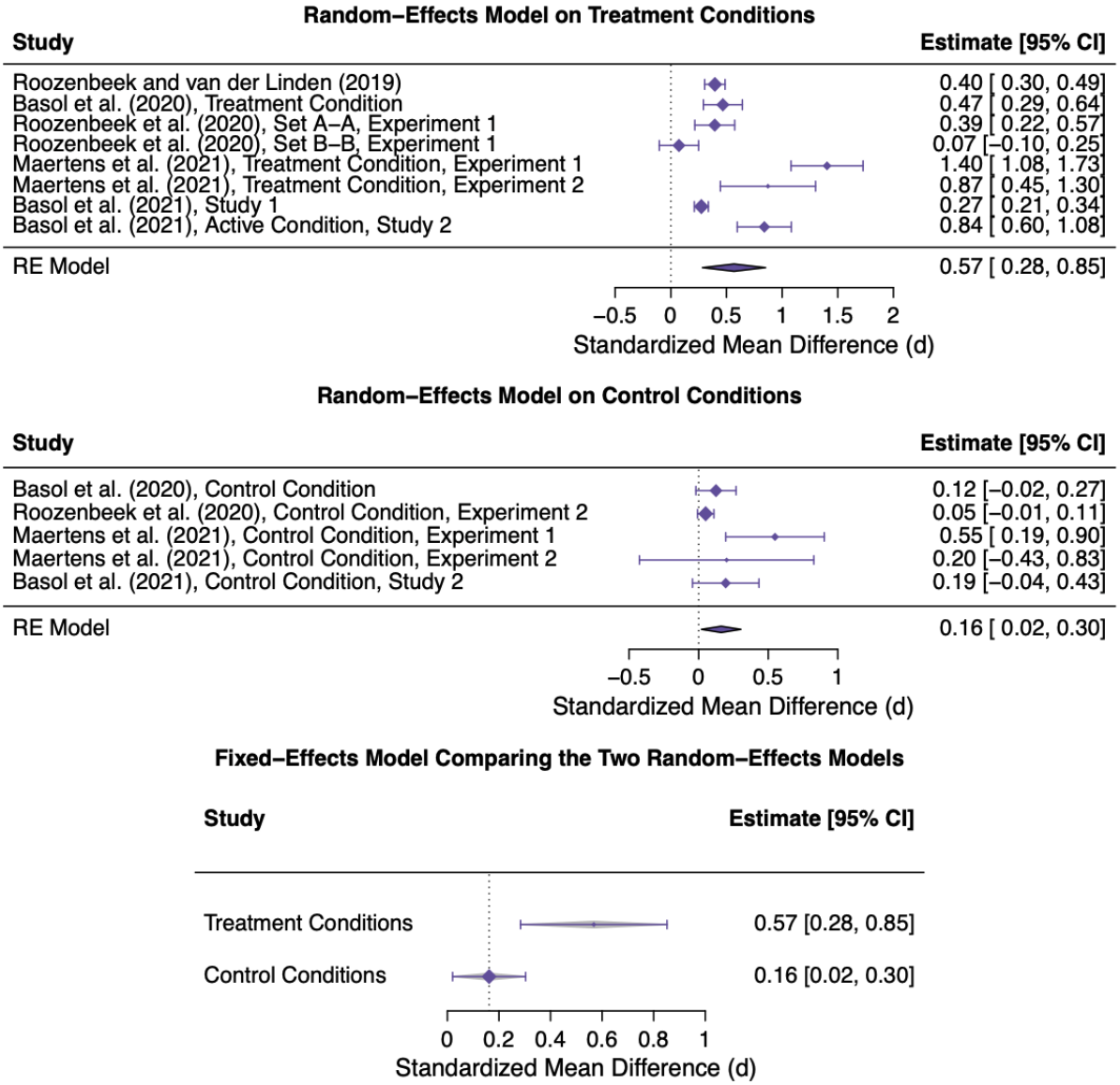
Note. RE = random effects. The error bars represent the standard errors. The size of the points represents the weight given to each study.

The results of our meta-analysis on the pre-post *B"D* effect sizes are shown in Table 3 and Figure 12. The random-effects model on the treatment conditions showed a significant increase in *B"D* values from the pre-test to the post-test, as did the random-effects model on the control conditions. However, the fixed-effects model comparing the two random-effects models showed a significant difference in pre-post *B"D* effect sizes between the treatment

conditions and the control conditions. Specifically, the increase in *B"D* values from the pre-test to the post-test was significantly greater in the treatment conditions compared to the control conditions. In other words, there was overall evidence for intervention-based effects on response bias, whereby the interventions made participants respond significantly more conservatively.

Figure 12

The Two Random-Effects Models and the Fixed-Effects Model We Fitted on the B"D Pre-Post Effect Sizes



Note. RE = random effects. The error bars represent the standard errors. The size of the points represents the weight given to each study.

General Discussion

The main goal of this paper was to explore whether gamified fake news interventions specifically affect only the targeted behavior (i.e., belief in fake news) or have more general effects that include the targeted behavior as well as other behaviors (e.g., belief in true

news). We achieved this goal using ROC analysis, a method commonly used with SDT that is ideally suited for determining the generality of interventions. Specifically, ROC analysis allows researchers to separate discrimination (i.e., the ability to distinguish between true and fake news), which is also referred to as news veracity discernment, from response bias (i.e., the tendency to rate news items as true or fake regardless of their objective veracity; Batailler et al., 2022). If the gamified intervention affects only the target behavior, this will result in an increase in discrimination because the FARs (derived from fake news ratings) would decrease whereas the HRs (derived from true news ratings) would remain static (or possibly increase). If the gamified intervention has more general effects, this will result in an effect on response bias because both the FARs and HRs would decrease. Despite this, to the best of our knowledge, ROC analysis has only been used once before with research on fake news (Modirrousta-Galian et al., in press). Instead, most studies have analyzed mean ratings, which are not ideally suited for separating discrimination and response bias (more on this later).

Consequently, we used ROC analysis to reanalyze the results from five different studies that used mean reliability or manipulativeness ratings to assess the effectiveness of Bad News and Go Viral!, respectively, two notable gamified inoculation interventions (Basol et al., 2020; Basol et al., 2021; Maertens et al., 2021; Roozenbeek et al., 2020; Roozenbeek & van der Linden, 2019). Table 4 shows a summary of the conclusions the authors drew from their data in those papers along with the conclusions we drew from reanalyzing the same data using ROC analysis. Overall, the authors of these prior studies concluded that Bad News and Go Viral! are effective in improving people's ability to detect false or misleading online content. Conversely, our reanalysis suggested that the interventions merely make people respond more conservatively, with little or no effect on discernment per se.

Table 4

A Summary of the Conclusions Drawn from the Reanalyzed Papers That Used Mean Ratings and Our Reanalysis That Used ROC Analysis

Intervention and Study	Conclusions from analyzing mean ratings	Conclusions from ROC analysis
Bad News		
Roozenbeek and van der Linden (2019)	Bad News improves people's ability to spot online misinformation.	Bad News caused a slight improvement in news veracity discernment, but this result can be attributed to differences in ambiguity between the true and fake news items.
Basol et al. (2020)	Bad News improves people's ability to spot online misinformation.	Bad News causes more conservative responding but not an improvement in news veracity discernment.
Roozenbeek et al. (2020)	Bad News improves people's ability to spot online misinformation, but item effects are present when using the same items in the pre-test and the post-test.	Bad News causes more conservative responding but not an improvement in news veracity discernment, and more conservative responding is more likely to occur when the news items used to assess mean ratings are ambiguous.
Maertens et al. (2021)	Bad News improves people's ability to spot online misinformation immediately after playing, but this effect decays and becomes negligible after 9 weeks unless participants undergo repeated testing.	Bad News causes more conservative responding but not an improvement in news veracity discernment. The increased conservative responding disappears after 9 weeks unless participants undergo repeated testing.
Go Viral!		
Basol et al. (2021)	Go Viral! improves people's ability to spot online misinformation but worsens their ability to spot true information immediately after playing. The effect on true news dissipates after 1 week, whereas the effect on fake news does not.	Go Viral! causes more conservative responding immediately after playing, but its impact on news veracity discernment is unclear; it was either ambiguous or improved after 1 week, but this latter result can be attributed to item differences as it also occurred to participants who played Tetris. Therefore, further research is needed to clarify this effect.

Note. ROC = Receiver operating characteristic. The list of conclusions drawn from analyzing mean ratings reported in the table are not exhaustive. That is, not all the original conclusions were reported, but rather just the ones that can be compared to those from our reanalysis.

The only studies that produced evidence of an intervention-based improvement in news veracity discernment when reanalyzed were Roozenbeek and van der Linden (2019) and Basol et al.'s (2021) Study 2 in the 1-week follow-up test. The result from Roozenbeek and van der Linden can be attributed to differences in ambiguity between the true and fake news items used to obtain mean reliability ratings. Specifically, the true news items were evidently reliable, whereas the fake news items were, by contrast, ambiguous (see reanalysis of Roozenbeek & van der Linden, 2019 presented earlier for more information). Clearly, it is much less likely for a psychological intervention to impact participants' beliefs in news items that they are certain about compared to news items that they are uncertain about. Indeed, when the differences in ambiguity between true and fake news items are less pronounced, as was the case for all the other reanalyzed studies on Bad News, no intervention-based improvements to news veracity discernment were found.

The result from Basol et al.'s (2021) Study 2 can also be attributed to item effects. Specifically, the items used in the 1-week follow-up test were easier than the items used in the pre-test and the post-test (see reanalysis of Basol et al., 2021 presented earlier for more information). This explains why the same delayed improvement in news veracity discernment was found in control participants who played Tetris.

Other Cases of Response Bias and Data Analysis Problems in Psychology

In our view, our research is one of many examples in the history of psychology where an initial interpretation of some data has been undermined by concerns about response bias. An early example occurred with the "dirty word" studies on perceptual defense conducted in the 1940s and 50s (see Eriksen, 1954 for a review). The idea under consideration at the time was that the ego protects consciousness from perceiving anxiety provoking stimuli such as taboo words. For example, McGinnies (1949) found that,

compared to neutral words, a longer exposure duration was needed before participants reported seeing briefly presented taboo words. Simultaneously, however, participants' Galvanic Skin Response (GSR) was raised even for the unreported subliminal taboo words, suggesting that participants unconsciously perceived the word, but the content of that perception was denied entry into consciousness. However, Howes and Solomon (1950) offered an amusing alternative explanation to these findings from a male participant's perspective:

> But, Heavens, NO! The word is – *penis*! And there is that girl (not to mention your professor) hanging on every word! Suppose it really isn't *penis*, after all (one can't be sure about tenth-of-a-second flashes)—what *wouldn't* they think about you if, out of the clear blue sky, you should volunteer *that* word! (p. 233)

Thus, failure to report the taboo words might have been due to conservative reporting, not failure to consciously perceive the taboo words. At the same time, the elevated GSR might have been due to participants feeling pressured by the context to say taboo words aloud.

Similar concerns about report bias have been raised – but in reverse – with hypnosis and the cognitive interview (Fisher & Geiselman, 1992) in memory research. Both procedures have been promoted as providing victims and witnesses of crime access to memories that would otherwise be unrecallable. However, some have noted that hypnotized participants tends to report more information than control participants (e.g., Klatzky & Erdelyi, 1985). Similarly, because one of the four main techniques incorporated into the cognitive interview is to "report everything," the cognitive interview is also likely to liberalize interviewees' report criterion (see Memon & Higham, 1999). If the report criterion is made more liberal, it is likely that there will be increased reporting of both true *and* false memories. However, if only the true memories are counted, a procedure that merely causes more liberal reporting may seem like a memory enhancement technique.

ROC analysis like that we have recommended here has the potential to identify more nuanced response bias problems. Indeed, some theorists have gone so far as to suggest that unless ROC curves are used, it is impossible to measure memory efficacy in such

common paradigms as old/new recognition (see Brady et al., 2022). In the context of eyewitness memory, Mickes et al. (2012) introduced ROC analysis to examine the relative efficacy of sequential versus simultaneous lineup procedures. The received wisdom at the time was that sequential lineups were superior. However, these conclusions were partly due to reliance on a measure known as the *diagnosticity ratio*, the ratio of the HR (proportion of correct identifications when culprit present) and FAR (proportion of incorrect identifications when culprit absent; diagnosticity ratio = HR/FAR).

Steblay et al. (2011) conducted a meta-analysis on studies that mainly used this measure. They noted that although both the HR and FAR were less in the sequential (vs. simultaneous) lineups, the FAR reduction was greater, which produced a higher diagnosticity ratio. However, the diagnosticity ratio is not independent of response bias under most circumstances, so these differences might be due to more conservative responding with the sequential lineup procedure rather than a *sequential-superiority effect*. Indeed, when Mickes et al. (2012) analyzed lineup data with ROC analysis, the opposite conclusion was reached (i.e., simultaneous > sequential), a finding that has had a dramatic effect on how the police use and interpret lineups.

In addition to the effect that ROC analysis has had on memory research, Heit and Rotello (2014) used ROC analysis to demonstrate that belief bias from the reasoning literature (the tendency to find fallacious reasoning to be valid when the content of the example is believable) was a response bias effect rather than a demonstration of flawed reasoning. Also, Rotello et al. (2015) argued that ROC analysis is needed in other domains such as shooter bias research in social psychology (the tendency for white participants to shoot unarmed black suspects more than unarmed white suspects) and referrals for child maltreatment. Although signal detection theorizing has been applied to both of these domains (e.g., see Correll et al., 2002, and Mumpower & McClelland, 2014, respectively), neither have used ROC analysis.

The longstanding illusion of a sequential-superiority effect is not just attributable to use of the diagnosticity ratio. A problem that is coupled with use of this statistic is analyzing

arithmetic differences in mean HR and FAR scores between the conditions of interest to assess discrimination. This problem is present in research on fake news as well. For example, Roozenbeek and van der Linden (2019) noted that:

> Although statistically significant, there were no meaningful differences in the pre-scores and post-scores of the "real" control headlines… In contrast, there were both statistically significant and much larger differences in the pre-scores and post-scores for the fake tweets (pp. 5-7).

An obvious interpretation of these statements is that the authors are assuming that the bigger pre/post difference scores for FARs (vs. HRs) is evidence that Bad News is improving discernment. This conclusion is reminiscent of Steblay et al.'s (2011) assumption that the bigger FAR (vs HR) difference between sequential versus simultaneous lineups was evidence for the sequential superiority effect. The problem with interpreting data in this way is that it assumes a linear relationship between the HRs and FARs. However, this assumption is not met in the vast majority of empirical psychological research, including fake news research (see Figures 4, 5, 8, 9, and 10). If, instead, the relationship is nonlinear, producing bowed ROC curves rather than straight lines, then shifts in response bias *on their own* can also produce a pattern of data where FARs change more than the HRs in response to an intervention.

What Does "More Conservative Responding" Mean?

In most of the reanalyses we conducted, there was evidence of both a lower FAR (fake news) and a lower HR (true news) in the post-test following the gamified intervention compared to the pre-test. This lowering of both the HR and FAR could be seen on the ROC curves as the points shifting towards the lower-left portion of the curve. It was also reflected in the SDT measure *B"D*, which indicated more conservative responding (i.e., increased *B"D* values associated with several of the ROC points in the post-test compared to the pre-test). How should we interpret these changes to response bias?

There are two SDT models that can account for more conservative responding in the post-test, which are shown in Figure 13 (Witt et al., 2015). The criteria-shift account shown

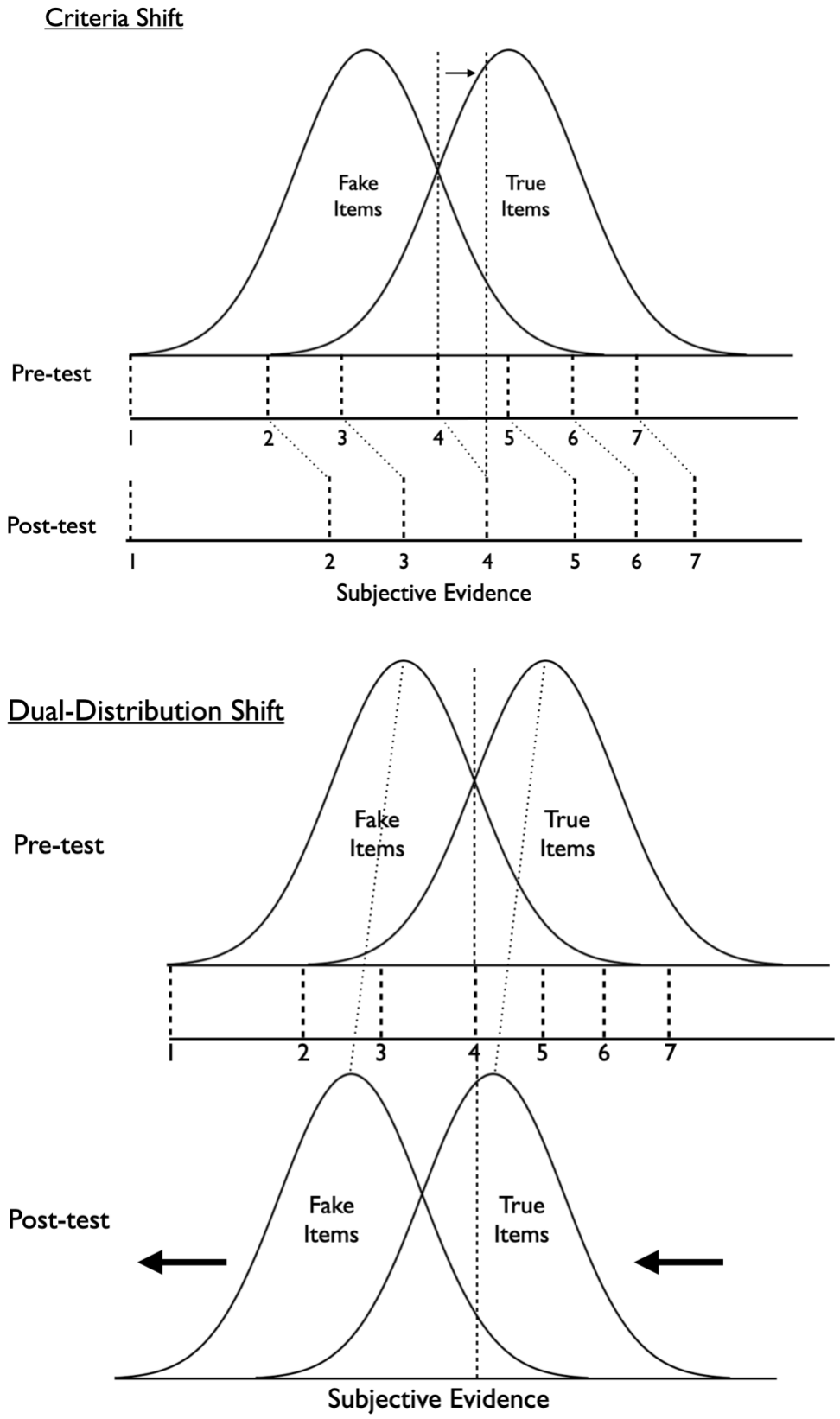
in the top panel of Figure 13 assumes that participants adopt more conservative criteria in the post-test compared to the pre-test. For convenience, Figure 13 shows a similar shift for each criterion. However, some criteria may shift more than others, while some may not shift at all, which is consistent with the *B"D* results from the reanalyses. For participants to adopt more conservative criteria in the post-test compared to the pre-test, it means that they need more subjective evidence of truth before assigning any scale value higher than 1.[7] This shift would have the effect of lowering both the HR and FAR at each of the scale points greater than 1, but not have any effect on discrimination.[8] In Figure 13, this decrease in both the HR and FAR is shown most clearly with scale point 4. For the pre-test, the 4 criterion is located at the intersection point of the fake and true news item distributions. However, at the post-test, assigning a 4 requires more subjective evidence, represented by the criterion for scale value 4 moving to the right of the intersection point. This shift means that fewer items are assigned 4 or higher, thereby lowering both the HR and FAR.

[7] As noted earlier, the criterion for "1" on the scale is maximally liberal and has a HR and FAR fixed at 1.0 because all items are assigned 1 or higher due to the nature of the task. Indeed, some may argue that 1 on the scale does not have a criterion associated with it at all, but we include it in Figure 13 to facilitate exposition.

[8] In principle, it is possible for the criteria to shift and for there also to be a change in discrimination. However, because our reanalyses did not reveal any clear evidence for cases where discrimination changed between the pre-test and the post-test (except for cases where there were item effects), neither model depicted in Figure 13 incorporates this feature.

Figure 13

Two Signal-Detection Models that Account for More Conservative Responding Following a Gamified Intervention.



The second potential account of more conservative responding in the post-test compared to the pre-test is shown in the bottom panel of Figure 13. The placement of the criteria and distributions for the pre-test in this model are the same as in the first model. However, to account for the more conservative responding in the post-test, this model does not assume that the criteria move up the decision axis, but rather that both the true and fake news item distributions shift down the decision axis, towards the left. In this model, which we have dubbed the *dual distribution shift*, the criteria are static; that is, there is no change in the amount of evidence needed to assign particular scale values. Instead, the intervention has led participants to perceive less evidence of truth in both true and fake news items.

Note that the HRs and FARs in the pre-test and post-test are identical between the two models. This equality is seen most readily by focusing on scale value 4 in Figure 13. Moreover, because the HRs and FARs are identical between the two models, so is the measure of bias, *B"D*. *B"D* is a function of the HR and FAR, and so it is only responsive to the magnitude of those values; it cannot "know" whether a change in the HR, FAR, or both, is due to criteria shifts or dual distribution shifts. Finally, it is worth noting that these two models represent the extreme cases. In reality, the data from the reanalyses could also have been produced by the interventions causing some combination of criteria and distribution shifts.

Without further research, it is difficult to distinguish between these two models. Nonetheless, it is worth considering the potential psychological mechanisms underpinning each model. The criteria shift model suggests that participants' assessment of the evidence for truth of the items does not change as a result of the intervention. A participant that assesses an item with *x* amount of evidence in the pre-test still assesses it as having *x* amount of evidence in the post-test. In other words, the intervention has not taught participants anything specific. Instead, participants have lowered the scale value they are willing to assign to an item with *x* amount of evidence.

There are at least two reasons why this type of scale recalibration might occur. First, the intervention may simply highlight the problem of fake news and cause a general increase

in participants' skepticism. For example, the intervention might activate a schema about most people being untrustworthy. In this scenario, the intervention has not made participants any better at identifying manipulative techniques in fake news. Rather, participants become skeptical of all news and are less willing to assign high reliability ratings or low manipulativeness ratings to any given news item.

A second possible reason for scale recalibration could result from methodological concerns. In most of the studies that we reanalyzed, the number of true news items in the pre-test was much lower than the number of fake news items.[9] Along with manipulating payoff matrices (i.e., the relative benefit vs. cost of hits vs. false alarms, respectively), varying the base rate probability of the signal trials (which are true news items in the current scenario) is a classical method of varying criterion placement in signal detection experiments (e.g., see Ratcliff et al., 1992; Rhodes & Jacoby, 2007). If there are very few signal trials, people are less willing to give high confidence "signal" responses, exactly the outcome we observed across multiple datasets. However, unless participants are explicitly told about the signal base rates a priori, it takes several trials for participants to learn that the correct response is nearly always "noise" ("fake"). Consequently, the likely result of a low proportion of signal (true news) trials in the pre-post design would be relatively unbiased responding on the pre-test as base rate learning is taking place, followed by higher criteria in the post-test.

Regarding the dual distribution shift account, there are also at least two different psychological mechanisms that might produce this outcome. The main difference between this account and the criteria shift account is that participants are assessing the evidence differently rather than merely recalibrating their use of the rating scale. Specifically, they are subjectively perceiving less evidence of truth, but this is occurring for both true and fake news items. Such an outcome could occur if the gamified intervention promoted correct identification of manipulative techniques in fake news, but participants were also incorrectly

[9] Note, however, that this explanation cannot account for the more conservative responding caused by playing GoViral! since the study examining its effectiveness (Basol et al., 2021) used an equal number of true and fake news items.

identifying these techniques in true news when they were not there. Again, the intervention causing activation of schemas such as “most people are untrustworthy” might play a role here. However, in this case, activation of this schema has changed the perception of the available evidence rather than how the rating scale is calibrated with respect to the subjective evidence.

An alternative psychological account of the dual distribution shift model is that some of the manipulative techniques the games seek to teach participants about are present in *both* fake news *and* true news. Consequently, if the intervention facilitates identification of these techniques, then it stands to reason that participants would perceive less evidence of truth (or more evidence of falsity) in both true and fake news items. Such a possibility is not without merit. Mainstream news outlets are under pressure to capture their audience’s interest and attention in the same way that generators of fake news are. In this vein, Hart et al. (2020) found that six reputable newspapers (i.e., *USA Today*, *The Washington Post*, *The Philadelphia Inquirer*, *The New York Times*, *The Los Angeles Times*, *The Minneapolis Star-Tribune*, and *The Atlanta Journal-Constitution*), five of which are rated as high and one of which is rated as moderate on factual reporting by the fact-checking website Media Bias/Fact Check (https://mediabiasfactcheck.com/), provided highly polarized news coverage on COVID-19. Furthermore, it is well established that news headlines, regardless of source, often use loaded words to fearmonger (Glassner, 2004) and evoke other emotions in readers (Clark, 2006). Polarization and provocative emotional content are two of the manipulative techniques taught by Bad News, and fearmongering is one of the manipulative techniques taught by Go Viral! (Basol et al., 2021; Roozenbeek & van der Linden, 2019). Therefore, Bad News and Go Viral! may effectively teach people how to detect manipulative strategies, but if they are present in both true and fake news, these interventions could cause people to consider all news as less reliable or more manipulative.

New Studies and the Up-to-Dateness of This Paper

We acknowledge that this reanalysis has the potential to be out of date the moment it is accepted for publication, especially due to the current popularity of gamified fake news

interventions. Indeed, after the conclusion of our literature review, two new papers on the effectiveness of Bad News were published that met all of our inclusion criteria, namely Roozenbeek et al. (2022) and Iyengar et al. (2022). Although we cannot continue amending this paper to include all of the newest experiments as they get published, we reanalyzed these two particular studies in an attempt to make our reanalysis as up to date as possible before it is published. We will not reanalyze any more papers henceforth. Overall, Roozenbeek et al. showed no meaningful improvement in discernment but a meaningful increase in conservative responding after playing Bad News (see Table S33 and Figure S1). In contrast, Iyengar et al. showed a meaningful increase in discernment after playing Bad News (see Table S33 and Figure S2).

We offer three different explanations for Iyengar et al.'s (2022) uncharacteristic result: (a) sample differences; Iyengar et al. recruited an Indian sample, whereas all the other reanalyzed studies recruited Western samples. Perhaps Bad News is only effective for non-WEIRD (western, educated, industrialized, rich, and democratic) samples; (b) design flaws; although the pre-test items were different from the post-test items, they were not counterbalanced, and the experiment lacked a control condition. These methodological issues introduce the potential confounds of sequence effects, order effects, and item effects; and (c) Iyengar et al. may simply be an outlier; out of the 12 reanalyzed treatment conditions, only one has shown such considerable improvements in discernment. Nevertheless, although including Roozenbeek et al. (2022) and Iyengar et al. in the meta-analysis changed the magnitude of effects, it did not change their statistical significance (see Table S28 and Figures S3 and S4). Therefore, the conclusions we made from our reanalysis remain the same.

Conclusion

The difference in the conclusions drawn from analyzing mean ratings and ROC analysis is arguably the most important finding from this paper. It demonstrates that by conflating discrimination and response bias, and thus providing a rather coarse and imprecise overview of the decision-making process, an intervention can misguidedly appear

effective in producing its intended effects. However, after separating these diverse influences, and thus providing a more detailed and accurate breakdown of the different mechanisms at play, the same intervention can be revealed to be not so effective in producing its intended effects. Consequently, using a statistical method that offers such nuanced insights is of vital importance. ROC analysis is perfectly suited for this purpose, and we therefore recommend its use for assessing the generality and hence efficacy of interventions that aim to improve fake news detection.

Constraints on Generality

Our findings suggest that Bad News and Go Viral! do not improve news veracity discernment. Instead, they both elicit more conservative responding. Given that this result has been observed across five papers amounting to a total of 13 experiments (eight treatment conditions and five control conditions) and 17,867 participants, we believe that it will be reproducible with participants from similar subject pools, specifically WEIRD samples. However, we do not have evidence that our findings will occur for non-WEIRD samples (see Iyengar et al., 2022). In most of the reanalyzed studies, the stimuli consisted of a limited number of true and fake news items, the latter of which were often created by the researchers. Therefore, we expect our results to generalize to situations in which participants rate similar item sets. Indeed, when participants were tested on a larger set of true and fake news items that had been posted on the internet in the past, Bad News had no effect at all; it did not improve news veracity discernment nor elicit more conservative responding (Modirrousta-Galian et al., in press). Finally, we do not have evidence that our findings will generalize to gamified fake news interventions other than Bad News and Go Viral!. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

Context

Gamified psychological interventions designed for improving people's ability to spot online misinformation have become popular. To determine their efficacy, we believe it is vital to assess their *generality* (their effect on both true and fake news) because some interventions may affect the intended behavior as well as unintended behaviors (reduced belief in true news). Considering that reduced belief in true news can potentially have devastating consequences (e.g., rejecting the scientific truth that vaccines are important for personal and global health), this distinction is critical. Receiver operating characteristic (ROC) analysis is ideally suited for determining the generality of interventions, as it allows for *discrimination* (ability to distinguish between true and fake news) and *response bias* (tendency to rate news items as true or fake regardless of their objective veracity) to be measured separately. Despite its usefulness, ROC analysis has scarcely been used in online misinformation research (although see Modirrousta-Galian et al., in press). Consequently, we filled this gap in the literature by using ROC analysis to reanalyze data from published papers on gamified inoculation interventions, which allowed us to determine their generality and thus effectiveness.

References

Adjin-Tettey, T. D. (2022). Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. *Cogent Arts & Humanities*, *9*(1), Article 2037229. https://doi.org/10.1080/23311983.2022.2037229

Afolabi, A. A., & Ilesanmi, O. S. (2021). Dealing with vaccine hesitancy in Africa: The prospective COVID-19 vaccine context. *The Pan African Medical Journal*, *38*, Article 3. https://doi.org/10.11604%2Fpamj.2021.38.3.27401

Aleci, C. (2021). *Measuring the soul*. EDP Sciences. https://doi.org/10.1051/978-2-7598-2518-9

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), Article eaay3539. https://doi.org/10.1126/sciadv.aay3539

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281–311. http://dx.doi.org/10.1080/03637751003758193

Basol, M., Roozenbeek, J., Berriche, M., Uenal, F., McClanahan, W. P., & van der Linden, S. (2021). Towards psychological herd immunity: Cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. *Big Data & Society*, *8*(1), 1–18. https://doi.org/10.1177%2F20539517211013868

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), Article 2. https://doi.org/10.5334/joc.91

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2022). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science*, *17*(1), 79–98. https://doi.org/10.1177/1745691620986135

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. T. (2022). Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-022-02179-w

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences*, *118*(5), Article e2020043118. https://doi.org/10.1073/pnas.2020043118

Calvillo, D. P., Rutchick, A. M., & Garcia, R. J. B. (2021). Individual differences in belief in fake news about election fraud after the 2020 U.S. election. *Behavioral Sciences*, *11*(12), 175. https://doi.org/10.3390/bs11120175

Clark, C. M. B. (2006). *Views in the news*. LED Edizioni Universitarie.

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior, 42*(4), 1073–1095. https://doi.org/10.1007/s11109-019-09533-0

Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, *83*(6), 1314–1329. https://doi.org/10.1037/0022-3514.83.6.1314

Corrigan, P. W. (2016). Lessons learned from unintended consequences about erasing the stigma of mental illness. *World Psychiatry*, *15*(1), 67–73. https://doi.org/10.1002%2Fwps.20295

de Gardelle, V., & Kouider, S. (2009). Cognitive theories of consciousness. In W. P. Banks (Ed.), *Encyclopedia of consciousness* (pp. 135–146). Elsevier. https://doi.org/10.1016/B978-012373873-8.00077-3

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, *44*(3), 837–845. https://doi.org/10.2307/2531595

Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General, 121*(3), 275–277. https://doi.org/10.1037/0096-3445.121.3.275

Donaldson, W., & Good, C. (1996). *A'r*: An estimate of area under isosensitivity curves. *Behavior Research Methods, Instruments & Computers, 28*(4), 590–597. https://doi.org/10.3758/BF03200547

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), 13–29. https://www.nature.com/articles/s44159-021-00006-y

Eriksen, C. W. (1954). The case for perceptual defense. *Psychological Review*, *61*(3), 175–182. https://doi.org/10.1037/h0058094

Fisher, R. P., & Geiselman, R. E. (1992). *Memory-enhancing techniques for investigative interviewing: The cognitive interview.* (pp. xi, 220). Charles C Thomas, Publisher.

Glassner, B. (2004). Narrative techniques of fear mongering. *Social Research: An International Quarterly*, *71*(4), 819–826. https://doi.org/10.1353/sor.2004.0001

Grace, L., & Hone, B. (2019). Factitious: large scale computer game to fight fake news and improve news literacy. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–8. https://doi.org/10.1145/3290607.3299046

Grady, R. H., Ditto, P. H., & Loftus, E. F. (2021). Nevertheless, partisanship persisted: Fake news warnings help briefly, but bias returns with time. *Cognitive Research: Principles and Implications*, *6*(1), Article 52. https://doi.org/10.1186/s41235-021-00315-z

Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2022). *How to think about whether misinformation interventions work*. PsyArxiv. https://doi.org/10.31234/osf.io/gv8qx

Hart, P. S., Chinn, S., & Soroka, S. (2020). Politicization and polarization in COVID-19 news coverage. *Science Communication*, *42*(5), 679–697. https://doi.org/10.1177%2F1075547020950735

Heeger, D. (1997, November 12). *Signal Detection Theory* [Teaching handout]. Department of Psychology, Stanford University. http://www.cns.nyu.edu/~david/handouts/sdt-advanced.pdf

Heit, E., & Rotello, C. M. (2014). Traditional difference-score analyses of reasoning are flawed. *Cognition*, *131*(1), 75–91. https://doi.org/10.1016/j.cognition.2013.12.003

Higham, P. A., & Higham, D. P. (2018). New improved gamma: Enhancing the accuracy of Goodman–Kruskal's gamma using ROC curves. *Behavior Research Methods*, *51*(1), 108–125. https://doi.org/10.3758/s13428-018-1125-5

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (2016). Internal mapping and its impact on measures of absolute and relative metacognitive accuracy. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 39–61). Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199336746.013.15

Howes, D. H., & Solomon, R. L. (1950). A note on McGinnies' 'Emotionality and perceptual defense.' *Psychological Review*, *57*(4), 229–234. https://doi.org/10.1037/h0060881

Huotari, K., & Hamari, J. (2016). A definition for gamification: Anchoring gamification in the service marketing literature. *Electronic Markets*, *27*(1), 21–31. https://doi.org/10.1007/s12525-015-0212-z

Iyengar, A., Gupta, P., & Priya, N. (2022). Inoculation against conspiracy theories: A consumer side approach to India's fake news problem. *Applied Cognitive Psychology*. Advance online publication. https://doi.org/10.1002/acp.3995

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford University Press.

Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Kwantlen Polytechnic University. https://doi.org/10.17605/OSF.IO/HF7DQ

Johnson, J. (2022). *Worldwide digital population as of April 2022*. Statista. https://www.statista.com/statistics/617136/digital-population-worldwide/

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1420–1436. https://doi.org/10.1037/0278-7393.20.6.1420

Kanozia, R., Kaur, S., & Arya, R. (2021). Infodemic during the COVID-19 lockdown in India. *Media Asia*, *48*(1), 58–66. https://doi.org/10.1080/01296612.2021.1881286

Klatzky, R. L., & Erdelyi, M. H. (1985). The response criterion problem in tests of hypnosis and memory. *International Journal of Clinical and Experimental Hypnosis*, *33*(3), 246–257. https://doi.org/10.1080/00207148508406653

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modelling: A practical course*. Cambridge University Press. https://psycnet.apa.org/doi/10.1017/CBO9781139087759

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177%2F1529100612451018

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://psycnet.apa.org/doi/10.1037/xap0000315

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/jto.0b013e3181ec173d

McGinnies, E. (1949). Emotionality and perceptual defense. *Psychological Review*, *56*(5), 244–251. https://doi.org/10.1037/h0056508

McGuire, W. J. (1964). Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (Vol. 1, pp. 191–229). Academic Press. https://doi.org/10.1016/S0065-2601(08)60052-0

Memon, A., & Higham, P. A. (1999). A review of the cognitive interview. *Psychology, Crime and Law*, *5*(1–2), 177–196. http://dx.doi.org/10.1080/10683169908415000

Messori, A., Damuzzo, V., Agnoletto, L., Leonardi, L., Chiumente, M., & Mengato, D. (2019). A model-independent method to determine restricted mean survival time in the analysis of survival curves. *SN Comprehensive Clinical Medicine*, *2*(1), 66–68. https://doi.org/10.1007/s42399-019-00199-7

Micallef, N., Avram, M., Menczer, F., & Patil, S. (2021). Fakey: a game intervention to improve news literacy on social media. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), Article 6. https://doi.org/10.1145/3449080

Mickes, L., Flowe, H. D., & Wixted, J. T. (2012). Receiver operating characteristic analysis of eyewitness memory: Comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, *18*(4), 361–376. https://doi.org/10.1037/a0030609

Modirrousta-Galian, A., & Higham, P. A. (2022, December 21). Gamified Inoculation Interventions Do Not Improve Discrimination Between True and Fake News: Reanalyzing Existing Research With Receiver Operating Characteristic Analysis. https://doi.org/10.17605/OSF.IO/85BE7

Modirrousta-Galian, A., Higham, P. A., & Seabrooke, T. (in press). Effects of inductive learning and gamification on news veracity discernment. *Journal of Experimental Psychology: Applied*.

Morschheuser, B., Hamari, J., & Koivisto, J. (2016). Gamification in crowdsourcing: a review. *Proceedings of the 49th Annual Hawaii International Conference on System Sciences*, 4375–4384. https://doi.org/10.1109/HICSS.2016.543

Mukhtar, S. (2021). Psychology and politics of COVID-19 misinfodemics: Why and how do people believe in misinfodemics? *International Sociology*, *36*(1), 111–123. https://doi.org/10.1177/0268580920948807

Mumpower, J. L., & McClelland, G. H. (2014). A signal detection theory analysis of racial and ethnic disproportionality in the referral and substantiation processes of the U.S. child welfare services system. *Judgment and Decision Making*, *9*(2), 114–128. https://journal.sjdm.org/13/13422/jdm13422.html

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Pennycook, G., & Rand, D. G. (2019). Cognitive reflection and the 2016 U.S. presidential election. *Personality and Social Psychology Bulletin*, *45*(2), 224–239. https://doi.org/10.1177/0146167218783192

Pertwee, E., Simas, C., & Larson, H. J. (2022). An epidemic of uncertainty: rumors, conspiracy theories and vaccine hesitancy. *Nature Medicine*, *28*(3), 456–459. https://doi.org/10.1038/s41591-022-01728-z

Phelan, J. C., Cruz-Rojas, R., & Reiff, M. (2011). Genes and stigma: The connection between perceived genetic etiology and attitudes and beliefs about mental illness. *Psychiatric Rehabilitation Skills*, *6*(2), 159–185. https://doi.org/10.1080/10973430208408431

Pollack, I., & Hsieh, R. (1969). Sampling variability of the area under the ROC-curve of d′e. *Psychological Bulletin*, *71*(3), 161–173. https://psycnet.apa.org/doi/10.1037/h0026862

Ratcliff, R., Sheu, C.-F., & Gronlund, S. D. (1992). Testing global memory models using ROC curves. *Psychological Review, 99*(3), 518–535. https://doi.org/10.1037/0033-295X.99.3.518

Read, J., & Harré, N. (2009). The role of biological and genetic causal beliefs in the stigmatisation of 'mental patients'. *Journal of Mental Health*, *10*(2), 223–235. https://doi.org/10.1080/09638230123129

Read, J. (2011). Why promoting biological ideology increases prejudice against people labelled "schizophrenic". *Australian Psychologist*, *42*(2), 118–128. https://doi.org/10.1080/00050060701280607

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(2), 305–320. https://doi.org/10.1037/0278-7393.33.2.305

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2020). Disentangling item and testing effects in inoculation research on online misinformation: Solomon revisited. *Educational and Psychological Measurement*, *81*(2), 340–362. https://doi.org/10.1177/0013164420940378

Roozenbeek, J., Traberg, C. S., & van der Linden, S. (2022) Technique-based inoculation against real-world misinformation. *Royal Society Open Science*, *9*(5), Article 211719. https://doi.org/10.1098/rsos.211719

Roozenbeek, J., & van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*, Article 65. https://doi.org/10.1057/s41599-019-0279-9

Rotello, C. M., Heit, E., & Dubé, C. (2015). When more data steer us wrong: Replications with the wrong dependent measure perpetuate erroneous conclusions. *Psychonomic Bulletin & Review*, *22*(4), 944–954. https://doi.org/10.3758/s13423-014-0759-2

Saleh, N. F., Roozenbeek, J., Makki, F. A., McClanahan, W. P., & 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

See, J. E., Warm, J. S., Dember, W. N., & Howe, S. R. (1997). Vigilance and signal detection theory: An empirical evaluation of five measures of response bias. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(1), 14–19. https://doi.org/10.1518/001872097778940704

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/bf03207704

Steblay, N. K., Dysart, J. E., & Wells, G. L. (2011). Seventy-two tests of the sequential lineup superiority effect: A meta-analysis and policy discussion. *Psychology, Public Policy, and Law*, *17*(1), 99–139. https://doi.org/10.1037/a0021650

Swift, J. (2012). *Political lying*. Bartleby. https://www.bartleby.com/209/633.html (Original work published 1710)

Tasche, D. (2008). Validation of internal rating systems and PD estimates. In G. Christodoulakis & S. Satchell (Eds.), *The analytics of risk model validation* (pp. 169–196). Elsevier. https://doi.org/10.1016/B978-075068158-2.50014-7

Tay, L. Q., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2021). A comparison of prebunking and debunking interventions for implied versus explicit misinformation. *British Journal of Psychology*, *113*(3), 591–607. https://doi.org/10.1111/bjop.12551

Urban, A., Moore, J., & Hewitt, C. (2019). Fake it to make it, media literacy, and persuasive design: Using the functional triad as a tool for investigating persuasive elements in a fake news simulator. *Proceedings of the Association for Information Science and Technology*, *55*(1), 915–916. https://doi.org/10.1002/pra2.2018.14505501174

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

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

Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal Detection Measures Cannot Distinguish Perceptual Biases from Response Biases. *Perception*, *44*(3), 289–300. https://doi.org/10.1068/p7908