The Long-Term Effectiveness of Inoculation Against Misinformation

An Integrated Theory of Memory, Threat, and Motivation

Rakoen Marieke Maertens

Downing College

Department of Psychology

University of Cambridge

November 2022

This thesis is submitted for the degree of

Doctor of Philosophy

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee.

Rakoen Marieke Maertens

November 2022

The Long-Term Effectiveness of Inoculation Against Misinformation

An Integrated Theory of Memory, Threat, and Motivation

Rakoen Marieke Maertens

Abstract

For over 60 years, inoculation theory has been a key framework to understand resistance to persuasion, yet many critical questions have remained unanswered. This dissertation aims to provide a theoretical and empirical understanding of how resistance to persuasion effects decay over time. In the context of resistance to persuasion by misinformation, I offer 10 empirical experiments that shed new light on this question, including several methodological innovations. In **Chapter 2**, I propose a new model that integrates memory theories with motivation theories on inoculation. In **Chapters 3–6**, I evaluate the long-term effectiveness of inoculation in message-based, gamified, and video-based inoculation interventions, unveiling the underlying mechanisms of decay. In **Chapter 7**, I address methodological issues, including the effects of repeated testing, and unstandardised items, and the development of a new misinformation susceptibility test. In summary, this thesis advances our understanding of the mechanisms of decay in resistance to persuasion, and sheds light on the role of and interplay between memory and motivation. The new memory-motivation model brings a significant advancement to the field, as it taps into the memory literature of forgetting—a domain in cognitive psychology—to shed new light on a concept in social psychology, and enables a new approach to modelling the longevity of inoculation effects. In addition, I offer novel insights into limitations with current methodological paradigms, and demonstrate how new standardised measurement tools can be developed to more accurately map inoculation effects in future research. Finally, I discuss how the findings of this dissertation can inform not only inoculation scholarship, but also intervention designers, evaluators, and policy makers, on how to address the problem of misinformation, and demonstrate how to extend the long-term effects of inoculation in applied interventions.

Acknowledgements (Part 1)

If there is one thing I have learned from pursuing a PhD at Cambridge, it is how important your social community and close friends are if you want to succeed and get the most out of the PhD experience. I have seen peers thrive and fall during their PhD. Thanks to the people listed in this section, I had the luck to have an incredibly rewarding, exciting, and intellectually stimulating PhD experience, as well as a warm community and home in Cambridge full of the most spectacular events and new friends for life.

In this first part of the acknowledgements, I would like to thank my academic community, starting with my supervisor, **Professor Sander van der Linden**, who was simply the most fantastic supervisor that I could imagine for a PhD. A true rising star in the field that lifts all around him up to amplify the effects together, as a team. He always trusted and believed in me, was always ready to make time, and gave me opportunities to lead ambitious projects with private and public sector partners that go far beyond what the typical PhD student can do. It has been a true privilege to work with him, and I will be forever thankful. I would also like to thank my advisor, **Dr Lee de-Wit**, who was always genuinely interested in how things were going and regularly probed how I was doing, and with whom I founded the Cambridge University Behavioural Insights Team.

My academic home I found in Cambridge Social Decision-Making Lab, that thrives in the spirit of collaboration, ambition, and peer-to-peer support. I want to thank all the Lab members for the great times and support. In particular I would like to thank **Dr Jon Roozenbeek**, with whom I've collaborated enthusiastically in many projects, and **Dr Claudia Schneider** and **Dr Cameron Brick**, with whom I have shared many exciting conversations and social events. I also would like to thank everyone in the Department with whom I did projects, discussed ideas, and shared PhD experiences. In particular I would like to thank **Emily**, **Corinne**, **Andrés**, **Leor**, **Sakshi**, **Maris**, **Tessa**, **Melisa**, **Ondřej**, **Ahmed**, **Kayla**, **Patrick**, **Fatih**, **James**, **Trisha**, **John**, **Mikey**, **Kristian**, **Cecilie**, **Karly** and **Steve**.

Finally, I would like to thank **Professor Josh Compton**, who is a true inspiration to the community of inoculation scholars, with an incredible knowledge of everything related to inoculation theory and a relentless and enthusiastic promotion of its further development.

Acknowledgements (Part 2)

In this second part of the acknowledgements, I would like to thank my social community. Cambridge students have the privilege to be part of a "College", in my case, "Downing College". I am incredibly grateful for the home that Downing has provided to me. In particular, I would like to give a big thank you to **Fritz**, **Slava**, and **Mark** for the fantastic and invaluable friendship since my very first arrival at the college and still going forward, as well as **Pierre**, **Joanna**, **Friedrich**, **Madeleine**, **Julian**, **Jacques**, **Jean**, **Anna**, **Elodie,** **Constantin**, **Dio**, **Rose**, **Dominic**, **Ryan**, **Iris**, **Pablo**, **Cameron**, **Faith**, **Geoffrey**, **Corbin**, and **Felix**. I would also like to thank the Chaplain, the **Revd Dr Keith Eyeons**, who has a wonderfully inspiring vision towards this world and organised the most spectacular evensongs. Just as important, is the community of friends outside of Downing, who have supported me throughout, starting with my incredible housemates at Parker Street. I would like to thank in particular **Raul**, **Caroline**, **Jivan**, **Olivier**, and **Oliver**, with whom I spent the most incredible of times. And then there are the friends I have met through Cambridge's rich social life. In particular I want to thank **Jamie**, **Owen**, **Wilhelm**, **Carl**, **Jenny**, **Eno**, **Edward**, **Richard**, **Steve**, **Tharpa**, **Jan**, **Sen**, **Chloé**, **Jon**, **Junaid**, **Julius**, **Malte**, **Lara**, **Katy**, **Felix**, and **Jonathan**. I also would like to thank my friends from Belgium, in particular **Olivier**, **William**, **Rien**, and **Robin**, who have been there for me since early in my undergraduate degree. Finally, there is a large international community that inspired me to keep travelling, and in particular I want to thank **Francesco**, **Sara**, **Michael Boris**, **Sasha**, and **Angel** for the many adventures abroad.

Then there is my family, and in particular **my parents**, who have always been the number one supporters in everything I do, giving me all the freedom, love, and trust I needed to venture out to England, always reminding me that there is a warm home to come back to. I also would like to thank **my godfather**, who has been an inspiration and a motivational force since before I started University, and helped make sure I reached all my academic potential.

Finally, I would like to thank one of the most wonderful people I have ever met, **Rue**, whom I had the pleasure to get to know and spend the most beautiful time with during the last two years. There are not enough words to describe how special she has made the past two years, and how much motivation, inspiration, and support she has given me every day. She made the past two years not only a great and exciting period, but the best time of my life so far.

Publications

The below list contains the citation details for the studies presented in the dissertation, providing the author order based on contributions and the publication status (i.e., published in a peer-reviewed academic journal, published as a preprint, or manuscript in preparation).

Study 1 – Journal Publication

Maertens, R., Anseel, F., & van der Linden, S. (2020). Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *Journal of Environmental Psychology*, *70*, Article 101455. https://doi.org/10.1016/j.jenvp.2020.101455

Study 2 – Manuscript in Preparation

Maertens, R., Simons, J., Roozenbeek, J., van der Linden, S. (2022). *The long-term effectiveness of consensus-based inoculation messages: Mechanisms*. Manuscript in preparation.

Study 3 – Journal Publication

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap000031

Study 4 – Manuscript in Preparation

Maertens, R., Simons, J., Roozenbeek, J., van der Linden, S. (2022). *The long-term effectiveness of gamified inoculation: Mechanisms*. Manuscript in preparation.

Study 5 – Manuscript in Preparation

Maertens, R., Roozenbeek, J., Lewandowsky, S., Maturo, V., Xu, R., Goldberg, B., van der Linden, S. (2022). *The long-term effectiveness of video-based inoculation: Three longitudinal experiments*. Manuscript in preparation.

Study 6 – Journal Publication

Roozenbeek, J., **Maertens, R.**, McClanahan, W., & van der Linden, S. (2021). 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

Study 7 – Preprint Publication

Maertens, R.*, Götz, F. M.*, Schneider, C. R., Roozenbeek, J., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., & van der Linden, S. (2022). *The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment*. PsyArXiv. https://doi.org/10.31234/osf.io/gk68h

Table of Contents

**The Long-Term Effectiveness of Inoculation Against Misinformation:
An Integrated Theory of Memory, Threat, and Motivation**

Chapter 1

General Introduction

1.1 The Problem of Misinformation

From global warming to the coronavirus disease 2019 (COVID-19), the world is facing major challenges that require a well-informed response. Yet, climate change deniers are spreading doubt-sowing messages on the scientific consensus (Cook et al., 2013, 2016; Oreskes, 2004), which seep into the processes of policymakers and hinder public action (Lewandowsky, Ecker, et al., 2017; Lewandowsky et al., 2015, 2019). Conspiracy documentaries discounting health advice on COVID-19 are going viral (Culliford, 2020), and the spread of misinformation has been linked to vaccine hesitancy (Loomba et al., 2021; Roozenbeek, Schneider, et al., 2020). Organised groups proclaim that 5G had a crucial role in creating or exaggerating COVID-19 symptoms, which led people to vandalise 5G masts (K. Chan et al., 2020). Disinformation can even lead to lynch mobs (Nugent, 2018).

An important debate within this domain discusses whether we have entered a post-truth era (Lewandowsky, 2020; Lewandowsky, Cook, et al., 2017; Lewandowsky, Ecker, et al., 2017), wherein partisanship, emotion, and perception are more important than objective reality. Research has found that misinformation spreads faster, more broadly, and deeper than real news (Vosoughi et al., 2018), especially when moral-emotional wording is used (Brady et al., 2017, 2020). Meanwhile, repeated exposure to a misinformation message can increase the perceived accuracy of such messages (known as the *illusory truth effect*), even when people do not believe it or when shown with a warning message (Hasher et al., 1977; Hassan & Barber, 2021; Pennycook et al., 2018). In addition, a long history of motivated reasoning research has shown that people give more weight to information congenial with their pre-existing attitudes and beliefs (Drummond & Fischhoff, 2017; Kahan et al., 2011, 2017;

Kunda, 1990; Lord et al., 1979). Partisanship may even result in ignoring the truth altogether, as seen with *expressive responding* (Schaffner & Luks, 2018). When political topics are discussed, it is possible that groups become polarized echo chambers (Barberá et al., 2015; Dunlap et al., 2016; Iyengar & Westwood, 2015; Motyl et al., 2014).[1] However, Pennycook et al. (2020) argue that people are still interested in the truth. Consistent with this idea, Pennycook and Rand (2019) found that analytical thinking is linked to a better ability to discern real news from fake news. Other researchers have found that actively open-minded (AOT) thinking has similar benefits and could decrease motivated reasoning (Bronstein et al., 2019; Carpenter et al., 2018; Stenhouse et al., 2018).

Over the past decade fact-checking websites have proliferated, with a major increase in interest since the Brexit referendum and the US presidential election in 2016 (Amazeen, 2020; Graves & Cherubini, 2016). Accordingly, a wide literature is dedicated to researching how to efficiently correct myths and depolarize groups (Lewandowsky, Cook, et al., 2020; Tay et al., 2022). But while fact-checking can indeed help decrease belief in a misinformation message (M.-P. S. Chan et al., 2017; T. Wood & Porter, 2019), there are caveats. First, the effectiveness can vary depending on how the debunking message has been crafted and how it is administered, with a risk of repeating the misinformation but not providing a coherent alternative (Lewandowsky et al., 2012). The correction can even *backfire* (i.e., have the opposite effect) when a group already has a pre-existing attitude or an opposing worldview that is incompatible with the correction (Nyhan & Reifler, 2010), although it has been shown that this rarely occurs and if it does, it typically is only present in highly polarised subgroups (see Nyhan et al., 2020; Swire-Thompson, DeGutis, et al., 2020; Swire-Thompson, Miklaucic, et al., 2022; T. Wood & Porter, 2019). Second, the correction benefit can decay over time while the myth holds a sustained influence beyond the correction, also called the

[1] Some scholars question the importance of echo chambers (see Garrett, 2017).

continued influence effect (CIE) of misinformation (Ecker et al., 2010, 2014; Johnson & Seifert, 1994; Lewandowsky et al., 2012, 2013; Swire, Berinsky, et al., 2017; Swire, Ecker, et al., 2017). Third, the success of the correction might partially depend on cognitive ability, in particular verbal ability, working memory, and episodic memory, which have been related to post-correction adjustment (Brydges et al., 2018; De keersmaecker & Roets, 2017; Sanderson et al., 2021).

In summary, the problem of misinformation is present and real (Lazer et al., 2018; Lewandowsky, Ecker, et al., 2017). Debunking may not always work and it is hard to keep up with fast-spreading misinformation (i.e., it is easier to produce misinformation than it is to fact-check every claim at scale). Therefore, recent literature reviews suggest a multi-layered defence system: in addition to the typical *de*bunking of misinformation, it would behove research to look into preventative approaches that focus on protecting people *before* exposure to the misinformation, also called *pre*bunking (Farrell et al., 2019; Roozenbeek & van der Linden, 2019a, 2019b; van der Linden, 2022; van der Linden & Roozenbeek, 2020).[2]

1.2 Resistance to Persuasion: Inoculation Theory

Persuasion research has traditionally focused on influence (Cialdini, 1993; Cialdini et al., 2006; Petty & Cacioppo, 1986), and much less on the opposite: fostering protection against influence (McGuire, 1970). Within the persuasion literature, *inoculation theory* has been described as *"the most consistent and reliable method for conferring resistance to persuasion"* (Miller et al., 2013, p. 127). Building on the biomedical analogy, William McGuire developed a theory focused on strengthening the cognitive immune system (McGuire, 1961a, 1961b, 1962, 1964, 1973; McGuire & Papageorgis, 1961, 1962). He

[2] Note that the terms "inoculation" and "prebunking" are often used interchangeably, but caution in doing so is warranted. While inoculation theory is typically applied in the context of preemptively protecting people against influence, it has become clear that inoculation can also be used as a debunking method (see Compton, 2020). I would therefore recommend to refer to "prebunking" and "debunking" when discussing the timing of the intervention (before or after exposure to misinformation), and to the intervention method (e.g., inoculation) irrespective of timing.

posited that the best way to build up cognitive immunity (i.e., "cognitive antibodies") against *persuasion attacks*, is to become familiar with the attitudinal challenge (i.e., "the virus") and its flaws (McGuire, 1970). This is done by combining two related components (Compton, 2013; see Figure 1.2.1), 1) an affective component: using a warning message to elicit a threat response (e.g., "beware, your attitude will be attacked"), and 2) a cognitive component: exposing people to a weakened dose of the attack (e.g., a counterattitudinal argument with the fallacies highlighted and alternatives provided). This enables people to both recognise and remember the attack as a threat, and to respond accordingly by preparing counterarguments (Eagly & Chaiken, 1993).

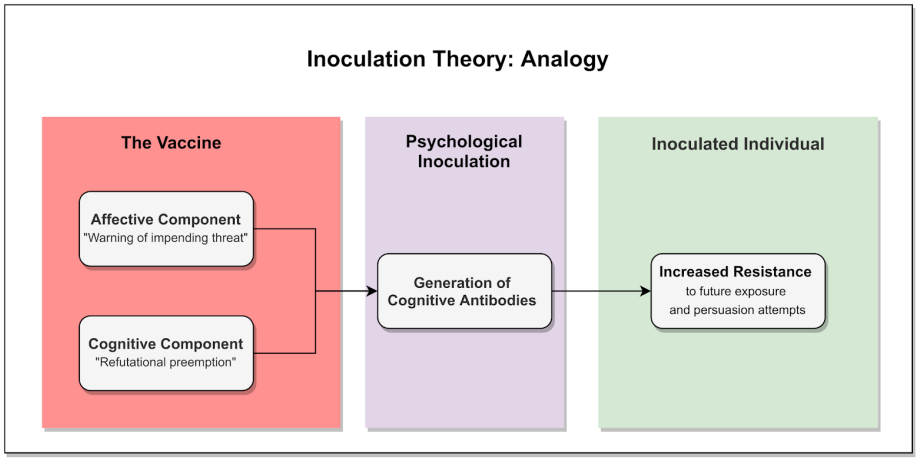

Figure 1.2.1. Basic components of inoculation theory in analogy to vaccination.

Inoculation theory has been in development for over 60 years and has proven to be robust both in isolated lab experiments on "germ-free" attitudes called *cultural truisms* (e.g., "brushing your teeth is good") and in more debated contexts such as health messaging (Compton & Pfau, 2005; Ivanov & Parrott, 2017; M. L. M. Wood, 2007). The theory recently has advanced to cover resistance to misinformation (Compton et al., 2021; Lewandowsky & van der Linden, 2021; Traberg et al., 2022), as seen in research on climate change misinformation (Cook et al., 2017; Maertens et al., 2020; van der Linden et al., 2017),

astroturfing (Zerback et al., 2021), and generalised inoculation against disinformation techniques (Roozenbeek & van der Linden, 2019a, 2019b; van der Linden & Roozenbeek, 2020). In a meta-analysis, a corrected average effect size of Cohen's $d = 0.43$ (95% CI: [0.39, 0.48]) was found when comparing inoculation treatments to no-treatment controls (Banas & Rains, 2010), which can be considered a large effect in the context of persuasion research (Funder & Ozer, 2019; Weber & Popova, 2012).

Recent developments have disambiguated three different parameters of inoculation theory. First, an attitude can either be protected before exposure to any counterarguments, making it a proactive or *prophylactic inoculation* (the classical form of inoculation), or corrected and then protected, making it a remedial or *therapeutic inoculation* (Compton, 2019).[3] Second, we have to define whether we want to inoculate against a specific misinformation message that is similar (*refutational-same*) to the inoculation message (a *narrow-spectrum vaccine*), or whether we want to provide blanket protection by providing a broad inoculation training (a *broad-spectrum vaccine*) that prepares for other (*refutational-different*) messages (McGuire, 1961a; McGuire & Papageorgis, 1961; Papageorgis & McGuire, 1961; Parker et al., 2016). Third, we should distinguish between *passive* forms of inoculation, where participants are presented counterarguments with little interaction, and *active* forms of inoculation, where people actively think about and interact with the materials to generate their own cognitive defences (McGuire, 1961a; Roozenbeek & van der Linden, 2019a, 2019b; van der Linden & Roozenbeek, 2020).

While *"the inoculation analogy is admittedly clever and valid"*, Eagly and Chaiken (1993, p. 568) highlighted, *"many of the questions it raised ... remain unresolved."* One of the most critical theoretical and practical questions that remain is that there currently is no

[3] Note that the difference between *prophylactic* and *therapeutic* inoculation interventions mainly refers to differences in pre-existing attitudes individuals have (i.e., whether only resistance to persuasion is needed or a correction as well) and not to actual differences in the procedure. The same intervention can be *prophylactic* or *therapeutic* depending on the participants and the content covered, and therefore not always clearly separable.

theoretical framework that can explain the long-term effectiveness and decay processes of inoculation interventions, nor has any attempt at predicting the decay curve succeeded (Banas & Rains, 2010).

1.3 Decay of the Inoculation Effect

Recent research has found that although the inoculation effect has been replicated many times, it has also repeatedly been shown that the inoculation effect decays over time (Banas & Rains, 2010; Ivanov et al., 2018). However, when and how quickly the inoculation decay process takes place, is still under debate. Multiple theoretical explanations have been proposed to try to explain and predict this decay in effect, but none have passed the test of empirical scrutiny (Banas & Rains, 2010). McGuire (1962, 1964) originally hypothesised that the optimal inoculation effect would be found only after an incubation period of several days, giving time for the cognitive defences to optimise by processing the learned information. While some evidence was indeed found for improved counterarguing after a short delay (Freedman & Sears, 1965; Hass & Grady, 1975; Petty & Cacioppo, 1979), others have shown the opposite (Insko, 1967; Pfau et al., 1997; Pryor & Steinfatt, 1978). A proposed explanation lies in a trade-off between 1) motivation to bolster counterarguments and 2) a decay of this motivation over time (Insko, 1967, p. 316), which led to the formulation of a *curvilinear relationship* hypothesis of inoculation decay (Compton & Pfau, 2005). This curvilinear relationship should show an increase in resistance over a short period after the inoculation, reaching a peak after a moderate delay, and finally followed by a decrease as time progresses. However, the meta-analysis by Banas and Rains (2010) did not find evidence for an initial increase in the inoculation effect, but rather a period of stability for at least two weeks, followed by a decay in the effect.

Issue involvement (i.e., how much do you engage with the topic), *post-inoculation talk* (i.e., do you talk about the inoculation content), *attitude accessibility* (i.e., how often do you

think about your own attitude towards the topic), and *attitude certainty* (i.e., how certain are you that you are correct) have been identified as important factors for bolstering attitudes and motivating to defend them against later attacks (Dillingham & Ivanov, 2016; Ivanov et al., 2012; Pfau et al., 2003, 2004, 2005). However, these variables have not yet been systematically studied in relation to decay.

In conjunction with the search for the underlying *mechanisms* of decay, we need to gain a better empirical understanding of the inoculation effect decay function. Pfau et al. (2004) found that counterarguing, a measure of resistance, is still present six weeks after inoculation. Moreover, Pfau et al. (1992) found that some inoculation effects may sustain over a period of seven months. Although in many studies the inoculation effect has shown remarkable resilience, they also show a diminishing effect over time (Banas & Rains, 2010; Ivanov et al., 2018). Research indicates that this decay might be less fast than other methods such as narrative messaging (Niederdeppe et al., 2015) or consensus messaging (Maertens et al., 2020), but the meta-analysis by Banas and Rains (2010) points toward decay starting after two weeks. However, more recent research shows that the inoculation effect may decay completely after two weeks (Zerback et al., 2021), or, in a different paradigm, may stay intact up to six weeks (Ivanov et al., 2018). These studies indicate that although the inoculation effect may remain significant for up to seven months for some interventions, depending on the method used, a decay process is likely to kick in anywhere between immediately after the intervention and six weeks after the intervention. The speed and mechanism of this decay process has not yet been systematically explored.

To counter the decay of the inoculation effect, evidence has been found for the effectiveness of *booster treatments* (Ivanov et al., 2018). It is theorised that just as for biomedical inoculation, a regular “booster shot” may be needed to top up the cognitive immune system (McGuire, 1961a). Examples of booster messages that have been proposed

include a weakened attack message, a repetition of the inoculation procedure (in full or shortened form), or a new warning message to elicit a fresh sense of threat (Ivanov et al., 2018). While evidence on the effectiveness of booster treatments is mixed (Compton & Pfau, 2005; Ivanov et al., 2009; Pfau, 1992), the general conclusion is that boosters work when administered in the right form at the right time but that more research is needed to answer these questions more accurately (Ivanov, 2012; Ivanov et al., 2018; McGuire, 1961a; Pfau et al., 2004). As Ivanov et al. (2018, p. 661) stress, *"the book on boosters is not ready to be closed"* and we need to *"reignite the research interest in inoculation booster messages"*.

In summary, while in recent years more empirical research has been done on the decay boundaries of inoculation theory, researchers have failed to develop a coherent theoretical framework that can explain the decay in resistance to persuasion or how to prevent it. As the meta-analysis by Banas and Rains (2010, p. 303) highlighted: *"more research about the inoculation decay process is needed."*

Chapter 2

A Memory-Motivation Theory on Inoculation

2.1 Bridging Cognitive and Social Psychology

Noticing this major gap in *"the grandparent theory of resistance to attitude change"* (Eagly & Chaiken, 1993, p. 561), I set out to develop and test a new theory of inoculation longevity that can account for previous findings as well as provide accurate new predictions. A currently unexplored area is that both inoculation decay and booster shot principles could be related to the more general insights into memory and the importance of rehearsal (Ebbinghaus, 1885; Hardt et al., 2013; Murre & Dros, 2015). In this dissertation, I will investigate whether the resistance to persuasion decay process can be explained in terms of *forgetting*, rather than motivation, or in conjunction with motivation (i.e., motivation strengthening memory).

Pfau et al. (2005) took a first step in this direction. Based on network models of memory (Anderson, 1983; Forgas, 2001), they theorised that resistance to persuasion might nest itself in associative memory networks. They argued that an attitude can be represented as an associative memory network with cognitive and affective nodes. Further, based on Petty et al. (1994), they argued that a more dense network is more resistant to change (i.e., resistance to influence from an attitudinal attack). Using a manual concept mapping exercise (participants were asked to freely write down and connect concepts that relate to the topic) as a proxy for mental structures they found increases in relevant concepts and linkages between concepts after an inoculation message, which in turn was related to more resistance to persuasion attacks at a later date (Pfau et al., 2005). However, other than this paper, a memory theory of inoculation theory has never been properly investigated.

An intuitive question might be whether it is not obvious that memory has to be involved in inoculation interventions; after all, if information is not stored in any type of memory, it cannot have any influence. Nevertheless, despite it being straightforward and with the exception of some explorations in the work by Pfau et al. (2005), memory has never been part of any major theoretical development within the inoculation literature, nor has it been compared with other theorised mechanisms involved in inoculation effects. Despite its omnipresence in research on misinformation and especially in the cognitive psychology of debunking (e.g., Ecker et al., 2010, 2014; Johnson & Seifert, 1994; Lewandowsky et al., 2012, 2013; Sanderson et al., 2021; Swire, Berinsky, et al., 2017; Swire, Ecker, et al., 2017), it has never been systematically studied in inoculation research. Understanding the role of memory could therefore contribute not only to understanding and improving the longevity of inoculation interventions, but also the initial strength and functioning of the interventions. This missing element in inoculation research may be due to inoculation theory having its roots in social psychology (McGuire, 1961a, 1961b) and communication science (Compton & Pfau, 2005), and researchers not realising how inoculation theory could benefit from the vast, existing literature on the cognitive psychology of memory (see e.g., Murty & Dickerson, 2016; Sanderson et al., 2021). I therefore argue that we should bridge the gap between social and cognitive psychology to allow new insights to be gained.

2.2 The Memory-Motivation Model of Inoculation

Further investigating the memory approach, we can link various concepts of inoculation theory to a potential theory about forgetting. It has been theorised that *active* inoculation is more potent than *passive* inoculation, which is in line with the benefits found with active and experiential forms of learning (Basol et al., 2020; McGuire, 1961a; Michel et al., 2009; van der Linden & Roozenbeek, 2020). Similarly, advantages for refutational-*same* over refutational-*different* interventions could be explained by the strength of their link with

the relevant associative memory network. Further, the concept of *booster treatments* can be interpreted as the benefit of rehearsal during the memory strengthening process (Ebbinghaus, 1885; Murre & Dros, 2015). Finally, as associative memory networks are part of the long-term memory system (Collins & Loftus, 1975; Smith, 1998), decay could be seen through the lens of neural network simulations of memory networks (Hardt et al., 2013).[4]

These findings provide a strong base to explore a memory theory of inoculation. Traditionally, the most likely explanation of both the baseline effects and its longevity, was seen as the threatening component (e.g., the forewarning) of the inoculation treatment leading to an increase in motivation to defend oneself (Ivanov, 2012). However, the notion of motivation is also compatible with a memory theory. Namely, it is well-known in the memory literature that motivation improves learning and leads to stronger memories (for a recent review, see Murty & Dickerson, 2016). A memory account on inoculation can therefore be compatible with the traditional account, augmenting it by unveiling its underlying mechanisms.

Based on the above findings, I propose a memory-motivation model of inoculation (see Figure 2.2.1). This model explains how inoculation works by creating, linking, and activating memory networks, starting with the initial inoculation memory creation during the intervention (the *learning* process). The strength of this first memory is moderated by the type of inoculation intervention (active/passive, prophylactic/therapeutic, specific/broad) and the motivation that was present during the learning process, whether elicited by the inoculation intervention or pre-existing. Over time, *forgetting* takes place, unless the memory is repeatedly strengthened (Frankland & Bontempi, 2005).

[4] Some memory models show a phenomenon called *catastrophic interference* (Lewandowsky & Li, 1995), memory voiding caused by replacing essential nodes in memory networks (cf. full decay of the inoculation effect), but also *partial reinstatement* (Atkins & Murre, 1998), the potential for full recovery of a memory network when part of the forgotten memory is reinstated (cf. booster treatments). It is however unlikely that *catastrophic interference* takes place in inoculation interventions, as inoculation messages are typically presented in a one-off fashion with limited conflicting information.

Forgetting refers to memories losing strength over time (Ebbinghaus, 1885; Thorndike, 1913), irrespective of the mechanism. Two theories on the mechanisms of forgetting have been proposed: *interference* (Underwood, 1957) and *trace decay* (Berman et al., 2009).[5] Trace decay (also known as temporal decay) refers to the gradual, automatic, decay of memory traces as a function of time. Interference takes place when new information (that typically is similar or related) makes it harder to access or retrieve old information (Oberauer & Lewandowsky, 2008). However, over the past decades, researchers have converged towards interference as the most likely explanation for forgetting, as there is little substantial evidence for trace decay, and *"there has been a long-standing consensus that [trace] decay plays no role in forgetting over the long term"* (Brown & Lewandowsky, 2010, p. 51).

The proposed theory provides a solution to forgetting: memory strengthening or *relearning* (cf. booster shots) could prevent or slow down forgetting and protect against *interference* by strengthening the memory network (Ebbinghaus, 1885; Ivanov et al., 2018; McGuire, 1961a). Finally, the model takes into account the underlying mechanisms proposed to be involved in robust inoculation interventions. *Issue involvement* (Pfau et al., 2005), the involvement of the individual with the topic and how important they find it, and *post-inoculation talk* (Dillingham & Ivanov, 2016; Ivanov et al., 2012), whether people talk more about the topic and the content of the inoculation intervention after the inoculation has taken place, are presented as potential moderators of the memory strengthening. Further, I list *booster treatments* as enhancers of the strengthening process (cf. *rehearsal* in classical memory models). Finally, *attitude certainty*, how certain someone is about their expertise on the topic, *attitude accessibility*, how easily someone thinks about the topic, and *intervention recall*, the objective declarative memory recall, are modelled as outcome indicators of a

[5] Note that within this dissertation, when I refer to *decay*, I use the functional definition of decay: a decrease in effect over time. I do not refer to decay as a possible explanation of forgetting—unless specified as *trace decay*.

successful strengthening process (Pfau et al., 2003, 2004, 2005). A strong memory in turn is hypothesised to provide long-term resistance to persuasion (i.e., less decay).

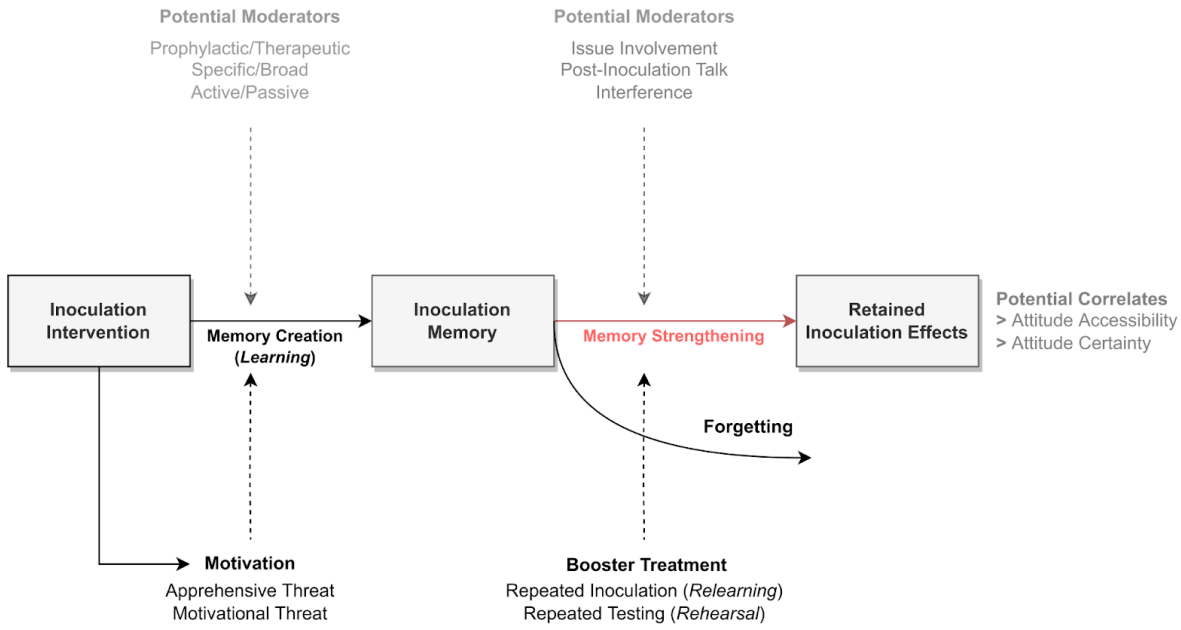

Figure 2.2.1. The proposed memory-motivation model of inoculation longevity.

In this dissertation, the focus will be on disentangling the influence of the two major components of the inoculation process, namely motivation and memory, and less on the potential moderators and correlates mentioned in the model, which can be explored in future research. Meanwhile, throughout the dissertation, I will always consider memory as long-term memory, and more specifically, explicit (declarative) memory (Camina & Güell, 2017). Explicit memory is consciously retrievable by the person and can include both specific learned knowledge (e.g., about the flaws in fake news), as well as a recall of the original event (e.g., the intervention experience). As it is expected that participants in an inoculation intervention are able to explain what they have learned and lucidly recall counter-arguments against misinformation, explicit memory will be my focus. It is feasible that other types of memory are involved, such as implicit memory (Camina & Güell, 2017; Tyng et al., 2017),

but this will not be explored in this dissertation. For the motivation term, when not otherwise specified, I refer to "motivational threat", or the self-reported motivation people have to defend themselves against misinformation, which is proposed to be introduced by the affective (threat) component of the inoculation intervention, and more predictive of inoculation effects than "apprehensive threat", which is the fearful or negative emotions towards misinformation generated by the intervention (Banas & Richards, 2017; Richards & Banas, 2018).

Further developing this approach, we can ask whether the inoculation decay function can be depicted as a forgetting function, and specifically, the Ebbinghaus *forgetting curve* (Ebbinghaus, 1885; Murre & Dros, 2015).[6] The forgetting curve is an exponential decay function with the steepness of decay being a function of memory strength and time, suggesting that a stronger memory—which can be attained by rehearsal or relearning (Karpicke & Roediger, 2008; Linton, 1975; Roediger & Karpicke, 2006a, 2006b)—will decay more slowly. Inoculation decay could be mapped vis-à-vis an Ebbinghaus decay function (see Figure 2.2.2 for an example of the theorised inoculation decay function in line with a forgetting curve), allowing for new predictions to be made, for example that with the right amount of booster sessions (i.e., sufficient memory strengthening), long-term psychological immunisation can be obtained.

[6] The forgetting curve was proposed by Hermann Ebbinghaus (1885) in his treatise *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie* [*Memory: A contribution to experimental psychology*]. 130 years later, the forgetting curve is still seen as robust and replicable (Murre & Dros, 2015).

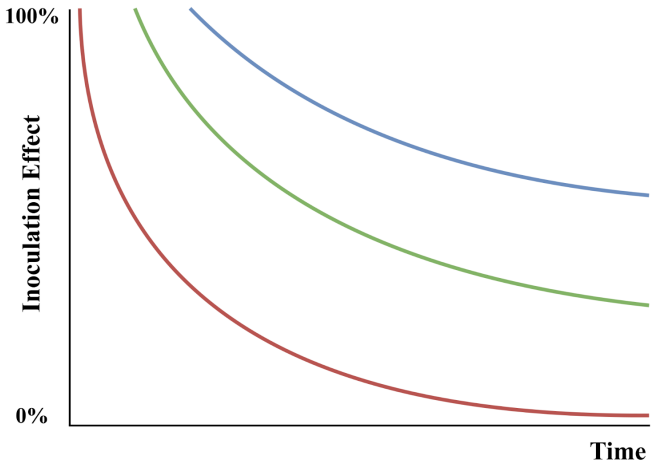

Figure 2.2.2. The theorised inoculation decay function (modelled as a forgetting curve) without rehearsal (red), with one rehearsal (green), and with two rehearsals (blue).

Within this dissertation, I theoretically and empirically investigate the validity of the proposed model. The overarching theoretical question that I ask is:

Theoretical Question

Can we explain the resistance to persuasion decay process using a memory-motivation theory of inoculation decay?

To gain insight into this question, I collect data over a series of longitudinal inoculation experiments, using a wide range of existing measures for memory strength, threat, and motivation (Pfau et al., 2005), and relate these with the decay process through longitudinal analyses based on the theoretical model in Figure 2.2.1. This will help to map the *mechanisms*. Meanwhile, I will further test the practical utility of the model with an overarching empirical question:

Empirical Question

What is the shape of the inoculation effect decay curve and can booster interventions remediate the decay?

Looking at the steepness of the decay by mapping a decay curve, we can establish whether the decay function follows an exponential decay pattern that resembles an Ebbinghaus formula, which includes a slower decay when memories are strengthened (e.g., through booster shots). To couple back to the theoretical question, I explore the decay and the potency of booster shots in relation to memory development, motivation, and threat.

Finally, in order to have the right toolkit to answer the theoretical and empirical questions, I ask a final methodological question:

Methodological Question

Is our methodological toolkit adequate to accurately measure the long-term effectiveness of inoculation effects?

Through two additional studies, I set out to explore potential methodological flaws in existing experiments, including testing effects, item effects, and the psychometric properties of our measurement instruments.

This dissertation aims to bring together two currently distinct branches of research, from multiple angles and through a variety of methodological approaches. The theoretical, empirical, and methodological frameworks of memory and resistance to persuasion research are both mature and robust, but connecting the two in a new and coherent memory-based inoculation paradigm, with an appropriate methodological backbone, has the potential to fill a major gap in persuasion scholarship as well as open up inoculation research to a wide range of new opportunities.

Chapter 3

Outline of the Dissertation and Inoculation Paradigms

Throughout the dissertation work I use three distinct experimental paradigms in order to test three main contexts (text-based, gamified, and video-based) of inoculation interventions and how the memory-motivation model of inoculation fits for each of them. First, in **Studies 1–2**, I look at inoculation decay in a passive narrow-spectrum therapeutic inoculation intervention (message-based inoculation, issue-focused). Next, in **Studies 3–5**, I look at inoculation decay in two broad-spectrum prophylactic inoculation interventions, one active (gamified inoculation, technique-focused), and one passive (video-based inoculation, technique-focused). Comparing the results of these opposite-end inoculation paradigms allows us to corroborate evidence for the validity and the generalisability of the proposed memory theory of inoculation, and will provide distinct insights for further theory development. Finally, in **Studies 6–7** I explore the methodological issues that were raised while working on these paradigms, and provide suggestions to improve further research.

In this chapter, I will first explain each paradigm concisely (more methodological details can be found in the methods section for each study), highlighting their basic stimuli and why they were chosen. I will then, in **Chapters 4–6**, establish for each paradigm the baseline effects, the decay curve, the effect of booster treatments, and the underlying mechanisms. These insights thereby provide data to respond to the theoretical and empirical research questions of the dissertation. Finally, in **Chapter 7**, I will explore item effects, testing effects, and the use of psychometrically validated measurement instruments.

Throughout the dissertation I strongly commit to reproducible open science (Nosek et al. 2015) and rigorous reporting standards (Appelbaum et al., 2018; Kazak, 2018; Levitt et al., 2018). For each study I therefore provide the raw and clean datasets, preregistration

documents, materials for replication (incl. survey files), power analyses, sample details, effect sizes, confidence intervals, and data cleaning and analysis code written in R.

To facilitate the finding of figures and tables presented in this dissertation, figures and tables are numbered in line with the chapter and section number. For example, the first figure or table in section 3.1 is numbered as 3.1**.1**. The third figure or table in section 4.2.4, is numbered as 4.2.4**.3**.

3.1 Message-Based Inoculation: Climate Change Paradigm

The first paradigm explores inoculation in the context of text-based climate change misinformation. Although 97% of climate scientists agree that human-caused global warming is happening, misinformation sowing doubt about the consensus influences society (Lewandowsky et al., 2015, 2019). Evidence suggests that debiasing public perception of the scientific consensus can lead to more support for collective action (van der Linden et al., 2015, 2019), but that this can be thwarted by misinformation (Cook et al., 2017; van der Linden et al., 2017). In the seminal study by van der Linden et al. (2017), participants were exposed to a consensus message (a pie chart depicting the scientific consensus, see Figure 3.1.1; van der Linden et al., 2014), an inoculation message (warning people not to be convinced by false petitions, and how this particular petition is flawed, see Figure 3.1.2), and a misinformation message (the misleading *Oregon Petition*, see Figure 3.1.1; Readfearn, 2016). They found that communicating the actual scientific consensus helps, as it helped to debias the *perceived scientific consensus* (i.e., people correct their belief about the scientific consensus), but that misinformation can neutralise all benefits. They also found that an inoculation message was able to protect this benefit by significantly reducing the impact of the misinformation message. The outcome variable measured is the *perceived scientific consensus on human-caused global warming*, on a slider scale from 0%–100%.

In this paradigm, the inoculation message is passive, issue-specific, and therapeutic. Passive, as participants read the messages without interacting with them (i.e., the experimenter provides the counter-arguments for the participant to read and remember). Specific, as it targets only the perceived scientific consensus, and presents a tailored inoculation message that includes a weakened version of the particular misinformation message (i.e., the message is focused on countering a specific piece of climate misinformation). And therapeutic, as on average people have (inaccurate) pre-existing attitudes regarding the scientific consensus on human-caused global warming (van der Linden et al., 2017).

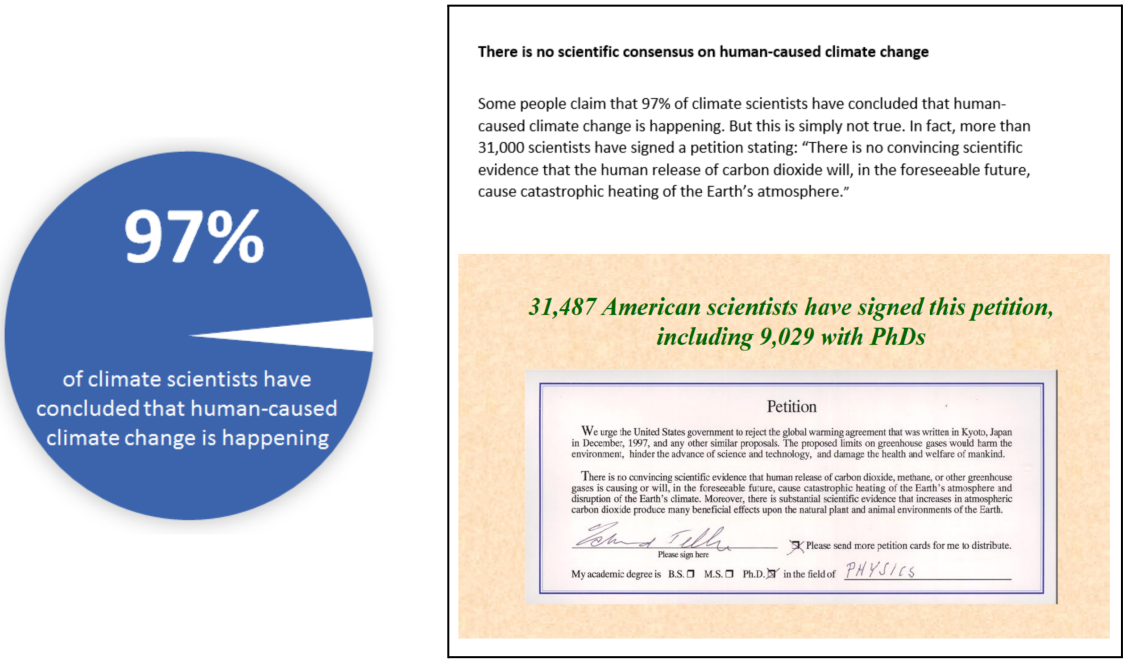

Figure 3.1.1. Consensus message (left) and misinformation message (right).

This paradigm was chosen as 1) it is a well-established inoculation paradigm (Cook et al., 2017; van der Linden et al., 2017; Williams & Bond, 2020), 2) the topic is relevant for both theory (i.e., inoculation using a debated and polarised issue) and society, and 3) it can provide novel insights into the validity of the memory theory of inoculation when using a passive, specific, and therapeutic inoculation intervention.

Nearly all climate scientists—97%—have concluded that human-caused climate change is happening. Some politically-motivated groups use misleading tactics to try to convince the public that there is a lot of disagreement among scientists. However, scientific research has found that among climate scientists "there is virtually no disagreement that humans are causing climate change".

One such politically motivated group claims to have collected signatures from over 31,000 "scientists" (including over 9,000 who hold Ph.D.'s) on a petition urging the U.S. government to reject any limits on greenhouse gas emissions because; "there is no convincing scientific evidence that human release of carbon dioxide, methane or other greenhouse gases is causing or will, in the foreseeable future, cause catastrophic heating of the Earth's atmosphere and disruption of Earth's climate." They claim that these signatures prove that there is no scientific consensus on human-caused climate change.

This may sound convincing at first. However, several independent investigations have concluded that the "Oregon Petition" is extremely misleading. For instance, many of the signatures on the petition are fake (for example, past signatories have included the long-deceased Charles Darwin, members of the Spice Girls, and fictional characters from Star Wars). Also, although 31,000 may seem like a large number, it actually represents less than 0.3% of all US science graduates (a tiny fraction). Further, nearly all of the legitimate signers have no expertise in climate science at all. In fact, less than 1% of those who signed the petition claim to have any background in Climate or Atmospheric Science. Simply calling yourself a "scientist" does not make someone an expert in climate science. By contrast, 97% of actual climate scientists, agree that human-caused climate change is happening.

Figure 3.1.2. The inoculation message used in the climate change studies.

3.2 Gamified Inoculation: Bad News Paradigm

The second paradigm uses an interactive online inoculation game called *Bad News*, developed by Roozenbeek and van der Linden (2019), in which people take the role of a fake news creator and spreader within a simulated Twitter-environment (see Figure 3.2.1 for a screenshot). In this inoculation intervention—which has already been played by over two million people and implemented in some school curricula in the United Kingdom and in Canada—the goal is to gain as many followers as possible by choosing and spreading misinformation messages while at the same time keeping your credibility sufficiently high. It includes the warning component of inoculation by showing the detrimental consequences misinformation can have (i.e., consequences of in-game actions) on topics that feel familiar (e.g., someone who gets fired from their job because of false accusations). This elicits a sense

of threat and motivation to resist similar persuasion attempts (i.e., the game warns people, and this motivates them to protect themselves against misinformation).

Unique is that people are exposed to "weakened doses" of broader misinformation *techniques* rather than specific issues, making it broad-spectrum. In other words, if people are inoculated against an entire technique (e.g., conspiracy theorisation), they should gain resistance to different variants of that technique (e.g., different conspiracy theories). It uses a framework of six influential misinformation techniques known as *DEPICT*: ***D**iscrediting opponents* (e.g., creating a cloud of doubt around your opponent)*, appealing to **E**motion* (e.g., the use of outrage or highly emotive language to manipulate people), ***P**olarizing audiences* (e.g., using hot-button issues to drive a wedge between two groups), ***I**mpersonation* (e.g., misusing the identity of politicians, experts, or celebrities online), *floating **C**onspiracy theories* (e.g., casting doubt on mainstream narratives by providing an attractive story in which a small sinister group of people is responsible for doing harm to many), and ***T**rolling* (e.g., eliciting reactions from people by provoking them online). See Roozenbeek and van der Linden (2019) and van der Linden and Roozenbeek (2020) for a detailed background and overview of these techniques. The active thinking, content creation, and choices people make for each misinformation technique serves as the cognitive component of the inoculation (i.e., through engaging with the weakened doses of misinformation, people generate preemptive refutations). This intervention serves as a broad-spectrum, active, and mainly prophylactic intervention. Broad, as it protects against a wide spectrum of different misinformation techniques (rather than specific misinformation messages). Active, as the intervention provides an experiential environment with interactive content. And mainly prophylactic, as the intervention is aimed at protecting existing attitudes from misinformation not seen before.[7]

[7] Although we cannot know the players' prior level of exposure to the misinformation tactics when they enter the game, the content of the game is fictional and therefore it can be assumed that there was no prior exposure to the specific content presented. However, participants might have seen or believed some misinformation using

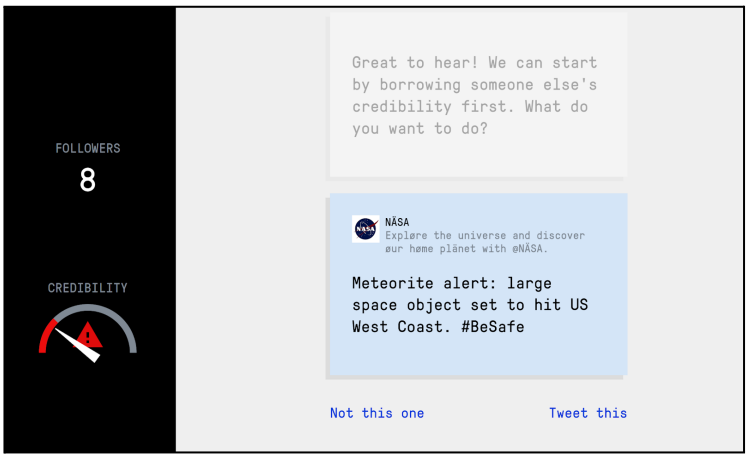

Figure 3.2.1. Screenshot of the *Bad News* game environment.

It has been shown that *Bad News* is effective at making people detect misinformation (Roozenbeek, Traberg, et al., 2022; Roozenbeek & van der Linden, 2019a)—replicated across cultures (Roozenbeek et al., 2020)—and at accurately increasing confidence in doing so (Basol et al., 2020). Resistance is measured by letting participants judge the reliability of real news items (that are neutral, non-misleading, and non-manipulative) and fake news items (that use one of the DEPICT techniques) on a 7-point Likert scale (see Figure 3.2.2 for an example), before and after the Bad News intervention.[8] However, as with the climate change paradigm, the long-term effectiveness of this paradigm has not been tested.

I chose the Bad News paradigm for the second range of studies as it 1) describes an applied, implementable, and widespread intervention, and 2) to test the memory-motivation model in a broad-spectrum, active, prophylactic inoculation intervention.

these techniques before (e.g., conspiracy theories), and therefore it could be argued that it may also function in parallel as a therapeutic intervention.

[8] Participants rate the fake news items as less reliable after the intervention, both compared to the pretest and compared to the control group (Roozenbeek & van der Linden, 2019a).

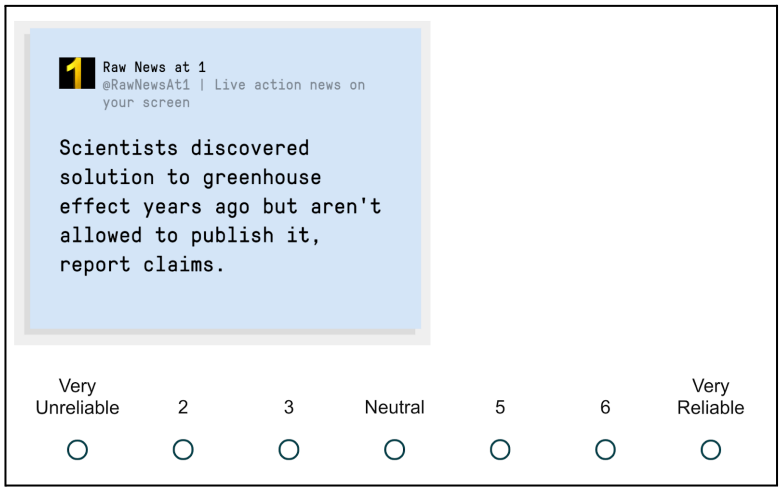

Figure 3.2.2. Example test item using the conspiracy technique.

3.3 Video-Based Inoculation: Emotional Language Paradigm

In a research project between Google Jigsaw, the University of Cambridge, and the University of Bristol, inoculation researchers designed and tested five short videos (~ 90 seconds), each of which "inoculates" viewers against a manipulation technique commonly encountered in online environments: emotional language (fearmongering), incoherence, false dichotomies, scapegoating, and ad hominem attacks.[9] In a first series of large randomized controlled trials (N = 5,416), the videos proved highly effective at 1) improving participants' ability to identify manipulation techniques in social media content; 2) increasing their confidence in their ability to spot such techniques; 3) strengthening their ability to discern trustworthy from untrustworthy content; and 4) improving the quality of their sharing decisions (Roozenbeek, van der Linden, et al., 2022). The videos are currently being rolled out as educational advertisements, and have been watched by over 5 million people. See Figure 3.3.1 for a screenshot of the emotional language video.

[9] See https://inoculation.science/inoculation-videos/ for an overview of the inoculation videos.

Figure 3.3.1. Screenshot of the emotional language inoculation video.

In this paradigm, participants watch a short inoculation video and subsequently rate 10 out of 20 possible social media posts (each participant receives the same 10 topics, but within each topic they have to rate either a manipulative or a non-manipulative variant of the item pair[10]) on a 1–7 scale (1—*strongly disagree*, 7—*strongly agree*), for each of the following dimensions: *Manipulativeness* ("This post is manipulative"), *Confidence* ("I am confident in my assessment of this post's manipulativeness"), *Trustworthiness* ("This post is trustworthy"), *Sharing Intention* ("I would share this post with people in my network"). Subsequently, a discriminative ability index for manipulativeness, trustworthiness, and sharing is calculated by subtracting the average scores of neutral posts from the average scores for the manipulative posts. For the confidence measure, results for manipulative and neutral posts are reported separately.

The intervention can be seen as a broad-spectrum, passive, and mainly prophylactic intervention. Broad, as it focuses on a general technique (e.g., *emotional language*) and thus protects against a wider range of potential misinformation messages. Passive, as people watch

[10] In addition to the 50% chance per topic of seeing the manipulative or non-manipulative variant, 5/10 topics contained only content extracted from actual social media sources (for both the manipulative and non-manipulative variants), and the other five topics used only fictive items specifically created for this experiment (for both the manipulative and non-manipulative variants). It is therefore possible that there is an imbalance in manipulative compared to non-manipulative items presented, but the ratio of real compared to created items is always balanced.

a video without any possibility to interact. And mainly prophylactic, as it aims to protect people's existing attitudes against messages and attacks they are not familiar with.[11]

I chose this final paradigm as it provides 1) a novel form of inoculation that is short and highly scalable, and 2) a test of the memory model for a broad-spectrum, passive (in contrast to the active *Bad News* intervention), prophylactic inoculation intervention, enabling the further evaluation of the generalisability of the model.

To close the outline, I would like to note that the purpose of the experiments in this dissertation is not primarily to test the interventions themselves or to comment on their effectiveness, but rather to use these different interventions to explore the important theoretical and practical questions about the decay process of the effects across different models and contexts of inoculation. By testing and comparing the findings across three different types of interventions, a more convincing and nuanced interpretation of the long-term effectiveness of inoculation can be attained, as well as a more thorough test of the proposed memory-motivation model.

3.4 Differences in Interventions and Differential Predictions

In the above sections I explained the different types of inoculation interventions that will be compared in the studies later in this dissertation. One such distinction made was the distinction between prophylactic and therapeutic interventions (Compton, 2020). In theory, this would mean that there are some interventions where there is an existing attitude we want to protect from change (prophylactic inoculation), and some interventions that are there to first correct an attitude and then protect it (therapeutic inoculation). Another dimension discussed is the difference between broad and specific interventions, where some interventions protect against one message (specific inoculation), and another type of

[11] Similar to the Bad News game intervention the content in the video is mostly fictional and therefore prior exposure should be limited, but some participants may be familiar with misinformation featuring the manipulative tactic (e.g., appeal to emotion) and may already have been influenced by it. It therefore may function as both prophylactic and therapeutic inoculation.

interventions protects against a wide range of messages (broad inoculation). However, in practice, the distinction between these concepts is not black and white, and therefore I would like to offer some nuance to the classifications from the previous sections. First, for the broad interventions, it may not be possible to properly distinguish between prophylactic and therapeutic interventions when using the classical definition. As we are protecting people against a wide range of misinformation, it could provide both prophylactic and therapeutic protection. In addition, in these experiments we cannot (cleanly) measure prior attitudes, as there is large variation in the amount of possible topics covered. Meanwhile, one could question if the climate change paradigm is truly therapeutic. In some countries, the perceived scientific consensus on human-caused global warming may be accurate, and therefore in some cases the focus may be on protecting the accurate attitude. Even in samples where this is not the case, there may be some participants for which the intervention works in a prophylactic way, as they may have accurate prior beliefs (this can be established with a pretest). Finally, it may be possible that in specific interventions, the effectiveness spills over to other, similar misinformation, using the same misinformation technique (e.g., misleading petitions)—making it more similar to what we would expect from a broad intervention. Taken together, the lines are blurry, and it may be better to consider them on a continuum, rather than a binary category.

The second question is whether we should expect differential findings in one paradigm or the other. The answer there is not clear-cut either. For example, some scholars have argued that *active* interventions facilitate resistance to persuasion and could improve the longevity of the inoculation effect, but evidence for this is mixed (Banas & Rains, 2010). Meanwhile, to date, no systematic research has been done to compare text-based, gamified, and video-based interventions with each other, especially not in relation to their long-term effectiveness, and therefore we cannot make any specific predictions on *how* they will be

different from each other. However, due to the substantial differences between the mediums, it is good to assume that there are *some* differences in decay between interventions. However, it is possible that these interventions only differ in the rate of decay, while keeping the same decay function, which would still allow for a generalisable model.

Chapter 4

Message-Based Inoculation

In two preregistered studies (**Study 1–2**), I explore the long-term effectiveness of inoculation and the memory-motivation model in a message-based paradigm. In the first study (**Study 1**), published in the *Journal of Environmental Psychology*, I investigated whether the inoculation effect decays over the course of one week without memory boosters, and compared the stability of the inoculation intervention to a consensus-only intervention. The study shows the feasibility and utility of a novel longitudinal inoculation design for message-based inoculation. It also shows that refutational-same (specific) interventions can elicit a strong initial memory even with limited interaction (as the intervention is passive) and using an already developed attitude (making it a therapeutic intervention), depicting results compatible with the memory theory of inoculation. For the published version of the paper, please see the full publication in the *Journal of Environmental Psychology* (https://doi.org/10.1016/j.jenvp.2020.101455). All supplementary materials, the preregistration, datasets, and analysis scripts in R are available on the OSF repository (https://doi.org/10.17605/OSF.IO/6BJSN).

In the second study (**Study 2**), I then extended these findings by providing a large-scale experiment with three time points and a range of questions exploring the underlying mechanisms of the effects. This study shows that memory is indeed one of the most important predictors of the inoculation effect, and that the inoculation effect can remain intact for at least 29 days without a booster intervention. Nevertheless, the study also shows that the effect decays over time, and that repeating the original intervention at a later time point can serve as a memory-strengthening booster. Finally, the study also shows evidence for the role of *both* motivation and memory, in line with the memory-motivation model proposed

in **Chapter 2**. All supplementary materials, including the preregistration, the data analysis scripts, the raw and clean datasets, and the survey files, can be found on the OSF repository at https://osf.io/9zxje/?view_only=44a8556694b54d09a2e2a9875071de2f.

Study 1

4.1 The Long-Term Effectiveness of Inoculation Against Misinformation: Message-Based Interventions[12]

4.1.1 Abstract

Despite the fact that there is a 97% consensus among climate scientists that humans are causing global warming, the spread of misinformation continues to undermine public support for climate action. Previous studies have found that resistance to misinformation can be induced by cognitively inoculating individuals against doubt-sowing about climate change. However, the long-term effectiveness of this approach is currently unknown. In a preregistered replication and extension experiment we combined a scientific consensus message with an inoculation treatment, and exposed participants to an influential misinformation message one week later. We explored 1) whether we can replicate the finding that inoculation is able to protect against a misinformation attack, and 2) whether or not the consensus and inoculation effects remain stable over the course of one week. Successfully replicating the effects of the original study, we found a strong initial consensus effect that is sensitive to doubt-sowing misinformation. Importantly, we also found that the consensus effect can be inoculated against misinformation. Extending the replication, we found that the consensus effect shows partial decay over time, while the inoculation effect remains stable for at least one week. We discuss the impact of our findings for inoculation theory, climate change psychology, and public policy.

[12] Study 1 has been published as *"Combatting Climate Change Misinformation: Evidence for Longevity of Inoculation and Consensus Messaging Effects"* in the Journal of Environmental Psychology (Maertens et al., 2020). The paper was written in collaboration with Professor Frederik Anseel (University of New South Wales) and Professor Sander van der Linden (University of Cambridge). I am the sole first author on this work.

4.1.2 Introduction

Climate change is one of the most pressing problems of our time (Allen et al., 2014), requiring large-scale collective action (van der Linden, Leiserowitz, et al., 2015). If no public action is taken, the continued rise in global temperatures could bring fundamental harm to human society and ecosystems. Numerous detrimental effects are already occurring, from extreme temperatures and floods to failed crop harvests and climate refugees (Biermann & Boas, 2010; Stocker et al., 2013). Indeed, many ecological systems are currently being threatened with destruction and biodiversity is falling drastically (Bridgewater et al., 2019).

Among climate scientists, a strong scientific consensus has been established on the fact that humans are causing global warming (Cook et al., 2016). Consensus research has replicated this finding in at least five high quality studies over the past years, estimating the consensus among scientists from a minimum of 91% to a maximum of 100% (Anderegg et al., 2010; Carlton et al., 2015; Cook et al., 2013, 2016; Doran & Zimmerman, 2009; Oreskes, 2004; Stenhouse et al., 2014; Verheggen et al., 2014). These studies further show a strong positive relationship between expertise in climate science and scientific consensus on human-caused climate change (Cook et al., 2016). However, most people are not climate science experts, and need to navigate facts through a cloud of (mis)information.

Consensus Messaging

Research has shown that communicating descriptive norms, such as the fact that 97% of scientists agree that humans are causing climate change, can positively influence belief in climate change and support for action, bridging the ideological divide (Goldberg et al., 2019; Kerr & Wilson, 2018; Lewandowsky et al., 2013; van der Linden et al., 2018; van der Linden, Maibach, et al., 2015, 2019). A recent dual-process model of attitudinal change on climate change, the Gateway Belief Model (GBM), shows how debiasing misperceptions about scientific norms can lead to higher perceived scientific consensus, which in turn serves

as a gateway belief with cascading effects on personal attitudes and support for collective action (van der Linden, Leiserowitz, et al., 2015, 2019). Although research on the benefits of communicating scientific consensus is now well-established (see van der Linden et al., 2019, for a recent review), research has also shown that when it is countered with an opposing (misinformation) message that contradicts the scientific consensus (e.g., a misleading petition claiming "there is no scientific consensus on climate change"), the positive effect of communicating the consensus is reduced or even completely neutralised (van der Linden et al., 2017).

Misinformation

Efforts to tackle the climate change problem have suffered from the influence of various forms of misinformation (Farrell, 2019; Lewandowsky et al., 2015; McCright et al., 2016; Oreskes & Conway, 2010). Among the many climate change misinformation techniques used, the most prevalent technique is sowing doubt about the scientific consensus (Oreskes & Conway, 2011). Through false-balance media coverage of the topic (e.g., 50%/50% instead of 97%/3%) or by using fake expert accounts (e.g., a professor in an unrelated field proclaiming to be a climate expert), perceptions of scientific consensus can be distorted (Cook et al., 2017, 2018; Koehler, 2016; Kortenkamp & Basten, 2015; Stocking et al., 2009).

At the same time, researchers find that online misinformation can spread faster and deeper than factual information, making it harder for scientific facts to reach the entire population (Lewandowsky et al., 2019; Scheufele & Krause, 2019; Vosoughi et al., 2018). Even when factual information corrects a myth, the initial belief (based on a falsehood) can still exert a continued influence (Ecker et al., 2014; Lewandowsky et al., 2012; Swire, Berinsky, et al., 2017; Swire, Ecker, et al., 2017). When the correction goes against the prevailing worldview of a polarized group, this correction may even backfire, though it

should be noted that the literature on backfire effects is increasingly debated (see Ecker et al., 2014; Ecker & Ang, 2019; Guess & Coppock, 2018; T. Wood & Porter, 2019). In short, although the 'debunking' approach to misinformation is often ineffective, it is still the most prevalent method used to tackle fake news (Graves & Cherubini, 2016). In contrast, it has been suggested that a more effective way of tackling misinformation is one based on building up resistance *before* any damage is done (Cook et al., 2017; van der Linden, 2019), an approach called 'prebunking'.

Inoculation Theory

Over the past 60 years, inoculation theory has emerged *"as the most consistent and reliable method for conferring resistance to persuasion"* (Miller et al., 2013, p. 127). Initially designed as a vaccine against brainwash by propaganda (McGuire, 1970), inoculation theory is now arguably the most established theory on resistance against persuasion (Eagly & Chaiken, 1993; Miller et al., 2013). Analogous to a biological vaccine, cognitive inoculation works by exposing people to severely weakened doses of misinformation ('the virus') to slowly build up cognitive resistance ('the antibodies') against misinformation (Compton, 2013; McGuire, 1962, 1964; McGuire & Papageorgis, 1961). The inoculation process consists of two components, namely; a) a forewarning of an attitudinal attack (the *affective* basis) and b) a process of refutational preemption (the *cognitive* basis). The underlying idea is that prior experience with persuasive arguments, combined with a warning of an upcoming threat, provides familiarity and alertness which can be used to disarm a later persuasive attack (Ivanov & Parrott, 2017). Meta-analyses have demonstrated that this method is generally effective, with effects (compared to no-treatment control groups) averaging on $d = 0.43$ (95% CI = [.39, .48]), which can be regarded as a strong effect size in persuasion research (Banas & Rains, 2010; Weber & Popova, 2012).

Initially, inoculation theory was developed in a 'germ-free' environment focusing on protection against attitudinal attacks that were relatively unquestioned (e.g., brushing your teeth). Recent research shows that inoculation is also effective within the context of a broad range of polarized issues, including climate change (Cook et al., 2017; van der Linden et al., 2017; M. L. M. Wood, 2007). Because people have differing prior attitudes on contested issues, conceptually, the inoculation approach is more therapeutic than truly preemptive. Accordingly, to better distinguish between inoculation interventions that try to "treat" an existing attitude and ones that purely protect, scholars have proposed to use the terms prophylactic and therapeutic inoculation (Compton, 2019). As people often have different underlying attitudes toward climate change, inoculation against climate change misinformation can be considered "therapeutic" in the sense that they boost immune response even among the already "afflicted."

In particular, van der Linden et al. (2017) designed an intervention which sought to combine climate change misinformation, consensus messaging, and inoculation. Results supported the effectiveness of inoculation against climate change misinformation. Specifically, the experiment had five conditions, including a) a facts-only condition where people were only exposed to a message about the near-unanimous scientific consensus on climate change, b) a misinformation-only condition where people were exposed to an influential misinformation campaign (the Oregon Global Warming Petition), which formed the basis of a viral story on social media that claimed that thousands of scientists have a signed a petition that global warming is a hoax (Readfearn, 2016), c) a false-balance condition where people were exposed to both the scientific consensus and the misinformation sequentially, and d) two inoculation conditions, both of which started with the scientific consensus message and then preemptively "vaccinated" people against the petition either with just a warning that some political actors try to deceive people on the issue of global

warming (the affective basis) or both a warning and a more detailed preemptive refutation of the petition (e.g., people were told beforehand that the petition has bogus signatories such as the *Spice Girls* and *Charles Darwin*). In the control group, people solved a neutral word puzzle. Attitudes toward the scientific consensus were assessed pre-and-post and the inoculation groups were exposed to the "full dose" (a screenshot of the debunked Oregon petition with a short description) at the end of the experiment. The authors found that, although misinformation had a significant negative effect ($d = 0.48$) by itself on perceived scientific consensus (and completely neutralised the positive effect of the scientific consensus in the false-balance condition), the inoculation conditions ($d = 0.33$ and $d = 0.75$) significantly protected attitudes toward the scientific consensus from harmful misinformation.

Cook et al. (2017) published a similar study using the same misinformation treatment with a focus on polarization based on differences in free-market support. They found that both the consensus message and the inoculation treatment were effective at lowering the negative and polarizing effects of misinformation, both for people low and high in free-market support. However, while both the study by van der Linden et al. (2017) and the study by Cook et al. (2017) provide valuable evidence on inoculation and consensus messaging, they do not provide any insights on the longevity of the potential inoculation effects, which is a crucial ingredient for the creation of interventions with long-lasting protection.

Long-Term Effectiveness

For its many years of development, there is still a lack of clarity about the long-term effectiveness of inoculation effects. In most studies, the delay between the inoculation and the persuasion attempt is only a matter of minutes or at most a few days (Banas & Rains, 2010). While most longitudinal studies do report decay to some extent, it is unclear what shape the decay function takes, and it has been understudied, in particular pertaining to topics that are

regularly debated in the media, such as climate change. Although some longitudinal studies have reported decay starting after two weeks, others have reported effectiveness of up to at least 44 days with booster messages (Banas & Rains, 2010; Ivanov et al., 2018; Pfau et al., 2006).

McGuire (1964) originally argued that a delay of a few days between inoculation and attack is needed for the effect to sink in maximally. While some evidence for the enhancement of resistance by delaying the attack has been found, there is more consistent evidence for the opposite conclusion: the inoculation effect dissipates over time (Banas & Rains, 2010; Ivanov, 2012). Potential explanations for the higher initial effect are the immediate sense of threat and memory salience (Ivanov & Parrott, 2017; Miller et al., 2013; Pryor & Steinfatt, 1978). The fresh sense of threat is elicited by the affective forewarning element within the inoculation treatment. This sense of threat provides heightened motivation to protect the attitude immediately after intervention. Over time, this warning could become less salient, and participants may be less vigilant to scrutinise incoming counterarguments. A different explanation is the decay of memory. Researchers have found that inoculation interventions strengthen associative memory networks, but this network could be subject to interference (Hardt et al., 2013; Lewandowsky & Li, 1995; Pfau et al., 1997, 2005). In their meta-analysis, Banas and Rains (2010) emphasise that *"more research about the inoculation decay process is needed"* (p. 303).

Replication Study

Most recently, Williams and Bond (2020) conducted a preregistered replication of van der Linden et al. (2017) using the same sampling platform (*MTurk*) and target population (US-only). The authors replicated many of the original findings, including the fact that the scientific consensus message and the inoculation intervention administered prior to the misinformation boosted attitudes toward the scientific consensus. Yet, there was one notable

exception: the authors did not find that the inoculation intervention counteracted misinformation to a greater extent than the scientific consensus message alone. As suggested by Williams and Bond (2020), potential explanations for this discrepancy include the presence of ceiling effects (most participants already scored near the maximum end of the scale on perceived consensus), or the fact that misinformation did not appear to lower perceived consensus in the false-balance or inoculation conditions. In fact, although the misinformation was effective by itself, its potency was much weaker ($d = 0.25$) than in the original study. Moreover, contrary to van der Linden et al (2017) where, in the false-balance condition, misinformation completely neutralised the consensus effect, in Williams and Bond (2020), perceived consensus (0%-100%) actually went up from 83.45 to 92.82 in the false-balance condition ($d = 0.55$, $p < 0.001$).

Present Study

In our present study, we set out two aims. First, we wanted to shed further light on the debate evoked by the replication study by Williams and Bond (2020). To do this, we decided to conduct our own preregistered replication and extension study of van der Linden et al. (2017). Second, we set out to contribute to the further expansion of inoculation theory by addressing the question of inoculation longevity in the context of a highly polarized and much discussed real-world issue: climate change. We asked the following key question: can we inoculate belief in the scientific consensus on climate change against a persuasive misinformation attack presented at a *later* date?

4.1.3 Methods

Design and Procedure

We investigated changes in *Perceived Scientific Consensus* (PSC) on human-caused climate change under different conditions that permit testing the long-term effectiveness of the inoculation effect. In contrast to van der Linden et al.'s (2017) original study where the

misinformation message was presented immediately after the intervention, we include a one-week interval between intervention and attack. Our study therefore consists of two phases, the first phase (including pretest T1 and posttest T2), and a second phase one week later (including posttest T3). The independent variables manipulated in our study were *exposure to the consensus message* (0, 1), *exposure to misinformation* (0, 1), *inoculation* (0, 1), and *test* (T1, T2, T3).

Following van der Linden et al. (2017), we designed our study in an additive format, where one new intervention is added per group, resulting in four different groups.[13] In the control group, participants were not exposed to anything, but did an unrelated word sorting task after pretest T1 instead to equalise the length of the task across conditions. In the consensus group, participants received the standard descriptive norm message about the scientific consensus after pretest T1. In the (false-)balance group, participants received the consensus treatment after pretest T1 as in the consensus group, but in addition to this a misinformation message one week later (just before posttest T3). Finally, in the inoculation group, the inoculation message was added immediately after the consensus message, and a misinformation message was presented one week later (just before posttest T3). Misinformation was thus not presented on the same day as the consensus or inoculation messages. This allowed us to eliminate potential short-term memory and demand effects, and helped us to test a decay hypothesis in which the benefit of consensus and inoculation messaging fully evaporates within one week. Further, to avoid demand effects, participants were told that they had randomly received the topic of 'climate change' out of 20 possible topics, and distractor questions were inserted after each intervention. See Figure 4.1.3.1 for an overview of the interventions in each group, and Figure 4.1.3.2 for a simplified graphical

[13] The original experiment had six groups, including two different inoculation conditions and a misinformation-only condition (van der Linden et al., 2017). Like Williams and Bond (2020), we omitted the partial inoculation group for simplicity. We pre-tested the misinformation-only condition (to make sure it is still effective) and therefore omitted it from the experimental design in the full study to preserve power for a longitudinal study.

flowchart of the full experiment procedure. A detailed overview of the exact steps from a participant perspective can be found in Supplementary Information S1 on the OSF repository for this study at https://doi.org/10.17605/OSF.IO/6BJSN. The experiment was approved by the University of Cambridge Psychology Research Ethics Committee (ref. PRE.2019.027).

		INTERVENTION		
		T2		T3
		Consensus Message	Inoculation	Misinformation
CONDITION	Control Group	x	x	x
	Consensus Group	✓	x	x
	Balanced Group	✓	x	✓
	Inoculation Group	✓	✓	✓

Figure 4.1.3.1. Overview of interventions per group.

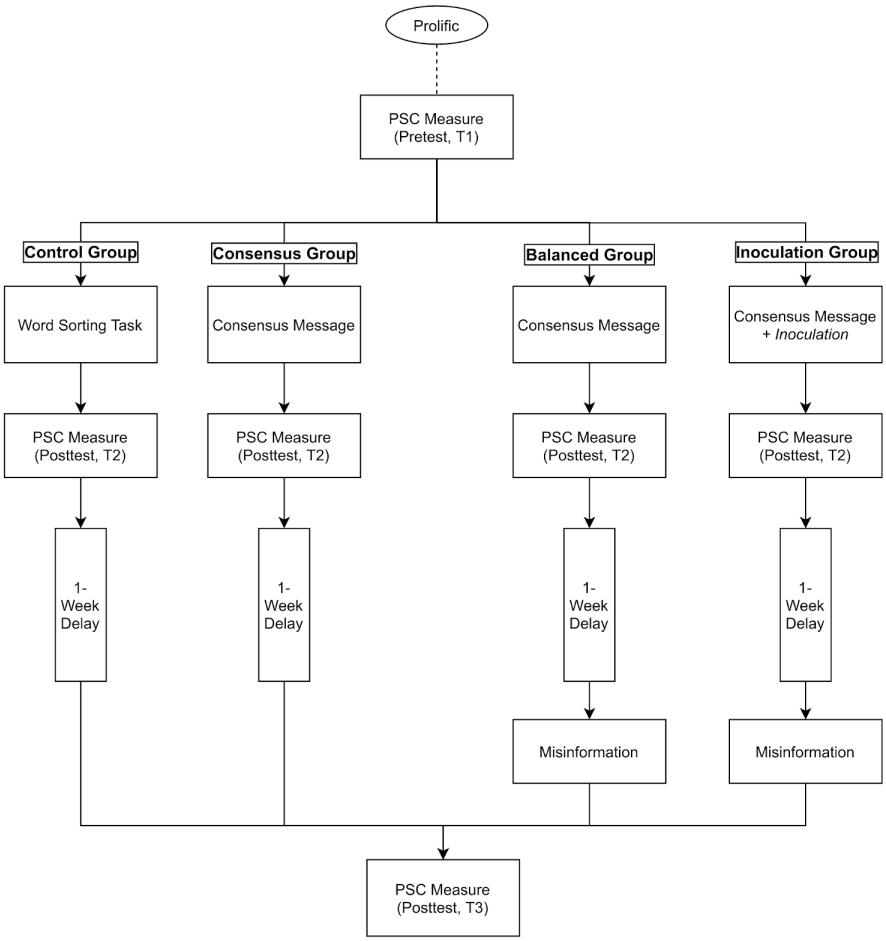

Figure 4.1.3.2. Flowchart of experimental procedure. PSC = Perceived Scientific Consensus.

Materials

Our core materials were adopted from van der Linden et al. (2017) and consist of a consensus message, a misinformation message, and an inoculation message. The consensus message simply informed participants that; "97% of climate scientists have concluded that human-caused climate change is happening."

The misinformation message used was a screenshot of the Oregon Global Warming Petition, which is a real but debunked petition summary that claims that "31,487 American scientists have signed a petition stating that human-caused climate change is not happening". This specific misinformation intervention is identical to those used by van der Linden et al. (2017), Cook et al. (2017), and Williams and Bond (2020). Van der Linden et al. (2017) initially selected this intervention as it was deemed most persuasive among a range of climate myths in a US nationally representative sample (N = 1,000).

Similarly, the inoculation method used is the exact same inoculation message used by van der Linden et al. (2017) specifically tailored to prebunk the Oregon Petition that includes both a forewarning (some politically-motivated groups use misleading tactics to try to convince the public that there is a lot of disagreement among scientists) and a detailed preemptive refutation (e.g., that the petition is debunked, contains bogus signatories, and that 31,000 may seem like a big number but only constitutes about 0.3% of all US science graduates).

A detailed overview of all materials can be found in Supplementary Information S2 on the OSF repository for this study at https://doi.org/10.17605/OSF.IO/6BJSN.

Measures

In random order, participants indicated their PSC, belief in climate change, belief in human causation, worry about climate change, and support for action. PSC was measured on a continuum (visual-analogue slider scale) ranging from 0-100 (M = 83.60, SD = 18.39), with

the question "*To the best of your knowledge, what percentage of climate scientists have concluded that human-caused climate change is happening?*", while the other questions were assessed using a 7-point Likert scale. Political ideology was measured on a 7-point Likert scale (M = 3.13, SD = 1.58), ranging from 1 (*very liberal*) to 7 (*very conservative*). A detailed overview of all variables and how they were measured can be found in Supplementary Information S3 on the OSF repository for this study at https://doi.org/10.17605/OSF.IO/6BJSN.

Hypotheses and Empirical Strategy

All hypotheses and analysis strategies were designed beforehand and preregistered.[14] All deviations from the preregistration are indicated in Supplementary Declaration S1 on the OSF repository for this study at https://doi.org/10.17605/OSF.IO/6BJSN.

As we wanted to look at how the consensus messaging effect decays over time, we first needed to replicate if we could indeed find a significant positive effect (van der Linden, Leiserowitz, et al., 2019). This led to our baseline hypothesis H_1.

[H_1] *In all intervention groups, consensus messaging has an initial positive effect on perceived scientific consensus.*

To measure whether misinformation neutralises the positive influence of the consensus effect and whether inoculation helps to protect this effect, the next question we asked is how the consensus message effect would evolve over the course of one week. Based on longitudinal studies comparing different messaging methods, we expected the consensus effect to retain its significance for at least one week (van der Linden, Leiserowitz, et al., 2019; van der Linden, Maibach, et al., 2019). This led to our second hypothesis (H_2).

[H_2] *The consensus message will retain its positive effect after one week.*

[14] https://aspredicted.org/bm6ny.pdf

To evaluate whether inoculation protects against misinformation that is presented one week later, the misinformation still needed to be effective at countering the consensus effect after a one-week delay. As the misinformation message was used in two different studies at two different time points, showing a similar negative effect, we also expected a significant negative effect in our study (Cook et al., 2017; van der Linden et al., 2017).

[H_3] *Misinformation presented one week after the consensus messaging treatment reverses the consensus messaging effect back to baseline.*

Finally, we wanted to investigate whether an inoculation message can indeed protect the newly changed belief against misinformation presented at a later point in time. Based on a meta-analysis of inoculation decay, we expected the inoculation effect to remain significant for at least one week (Banas & Rains, 2010).

[H_4] *When an inoculation treatment is added to the consensus message, the consensus effect remains significant, even after a misinformation attack one week later.*

To adhere to open science standards, we preregistered our study and provide full access to the anonymised dataset and all relevant materials (Nosek et al., 2015). Our preregistered analyses and any deviations from them are highlighted in Supplementary Declaration S1. Materials, datasets, and analysis scripts are publicly available on our OSF repository: https://osf.io/6bjsn/ (DOI: 10.17605/OSF.IO/6BJSN).

Sample

On the basis of the original climate change inoculation study where a specific inoculation effect size (compared to a no-treatment control) was found of $d = 0.75$ (van der Linden et al., 2017) and the more general meta-analysis effect size $d = 0.43$ (Banas & Rains,

2010), we hypothesised to find effect sizes of $d = 0.50$ or higher. We performed a power analysis to test our hypotheses with power = .95 and $\alpha = .05$, while accounting for a potential 30% attrition in participants between T2 and T3. On the basis of this analysis we recruited a total of 480 participants ($n = 120$ per group) through the online platform *Prolific Academic* (https://prolific.ac/). We limited the sample to participants of age 18 or above and US participants in order to match the design of the original study and Williams and Bond (2020) as closely as possible. Participants were told in the recruitment message that they would receive a total of 0.50 GBP if they participated in both Part 1 and Part 2 of the study. All participants gave informed consent before participation.

In the preregistration we stated that we would remove participants who either failed both attention checks or finished the first part of the study in less than one minute. All participants passed these checks, except for one who failed both attention checks; this case has been removed from the dataset. In the transition to Part 2 we lost 64 participants, which amounts to 13% attrition. As we had accounted for 30% attrition in our preregistered analyses, and to have comparable results across conditions, we use complete cases only for this study.[15] Neither political ideology nor pretest score could predict the attrition rate (see Supplementary Analysis S1 for a complete attrition analysis).

Within the complete-cases dataset ($N = 415$), 51% of participants were female with a median age of 34 ($M = 36.60$, $SD = 12.80$), they resided in 45 different US states, had a political ideology skewed towards left-wing (59% liberal, 19% conservative, $M = 3.13$, $SD = 1.58$), and a majority had a higher-education-level diploma (54%). As allocation to the different groups was random, the demographics are well balanced between the groups. All demographics data were self-reported through the survey. For a more extensive overview of the demographic variables, see Supplementary Information S4. The sample characteristics of

[15] The preregistration does not specify whether to use the full dataset or the complete cases dataset.

the studies by van der Linden et al. (2017), Williams and Bond (2020), and the current study are largely similar (see Supplementary Information S5).

4.1.4 Results

Pilot Study

To verify if the misinformation intervention is still effective two years after the initial study and in a US sample recruited via the *Prolific* platform instead of *Amazon MTurk*, we decided to run a pilot study with 80 participants. The only purpose was to replicate the negative effect of the misinformation intervention. Participants indicated their PSC (T1), were then exposed to the misinformation item (the Oregon Petition), and finally indicated their PSC once more (T2). According to the original study, we expected an effect size *d* close to .48 (van der Linden et al., 2017).

After performing a paired samples *t*-test, we found the expected effect size ($t(79) = -4.23$, $M_{diff} = -12.99$, 95% CI = [-19.10, -6.87], $p < .001$, $d = -0.47$), which indicated that this treatment is just as effective in decreasing PSC as in the original study. See Figure 4.1.4.1 for a visualisation of the T2-T1 difference.

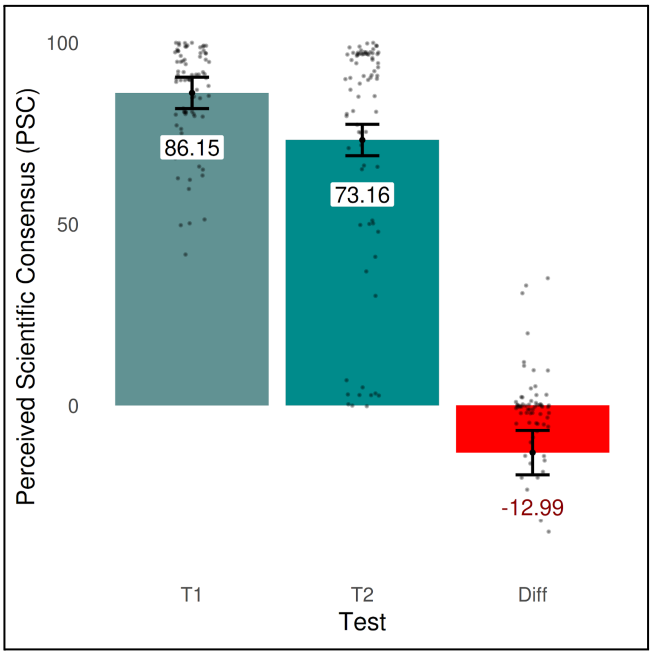

Figure 4.1.4.1. Means of *Perceived Scientific Consensus*, before (T1) and after (T2) exposure to misinformation. Error bars represent 95% confidence intervals. $N = 80$.

H_1: The Consensus Effect

Our first hypothesis stated that we should find a significant positive effect of exposure to the consensus message. To test our baseline hypothesis as preregistered, we used an ANCOVA with PSC at first posttest (T2) as the dependent variable, PSC at pretest (T1) as a covariate, and *group* as a between-factor. We found the expected significant effect of group ($F(3, 409) = 21.85$, $p < 0.001$, $\eta_p^2 = 0.14$).

In line with our preregistration, we compared the consensus group ($M_{diffT2T1} = 9.06$, $SD = 16.29$), the balanced group ($M_{diffT2T1} = 9.82$, $SD = 13.65$), and the inoculation group ($M_{diffT2T1} = 10.17$, $SD = 13.74$) using between-subjects tests to the control group ($M_{diffT2T1} = 0.68$, $SD = 3.19$), and expected a significant positive result for each comparison.[16] Compared to the control group, we find a significant effect for the consensus group ($t(113) = 5.19$, $M_{diff\text{-}in\text{-}diffs} = 8.38$, 95% CI [5.18, 11.57], $p < .001$, $d = 0.71$]), the balanced group ($t(111) =$

[16] The preregistration does not specify which test to use for the specific group comparisons. We chose to use difference scores in order for the results to be directly comparable to the approach of Williams and Bond (2020).

6.56, $M_{\text{diff-in-diffs}} = 9.14$, 95% CI [6.38, 11.91], $p < .001$, $d = 0.92$), and the inoculation group ($t(115) = 6.89$, $M_{\text{diff-in-diffs}} = 9.49$, 95% CI [6.76, 12.22], $p < .001$, $d = 0.95$). All findings are in line with our hypothesis.

Figure 4.1.4.2 depicts a bar plot visualising the within-group difference scores of the consensus treatment by group before exposure to misinformation, highlighting the similarity of the effect in each experimental group. A detailed overview of all pre, post, and within-subject difference scores can be found in Supplementary Analysis S2.

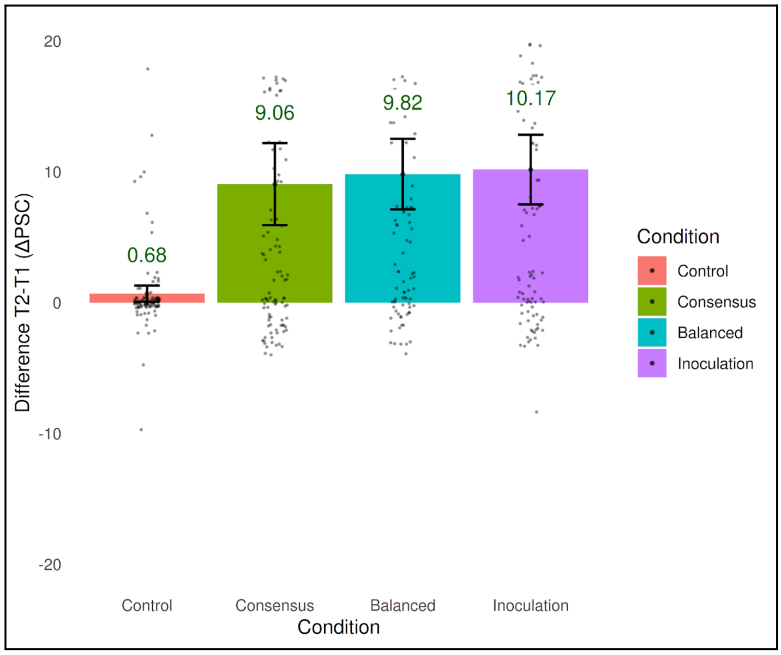

Figure 4.1.4.2. Comparison of T2-T1 *Perceived Scientific Consensus* difference scores (in percentage points) for each group. Error bars represent 95% confidence intervals. $N = 415$.

H2: Decay of the Consensus Effect

Our second hypothesis predicted that the consensus effect would still be found after a one-week delay. To test the second hypothesis as preregistered, we use an ANCOVA with the final PSC (T3) as dependent variable, *group* as between-factor, and T1 PSC as covariate. The days between T2 and T3 approximately resembled a one-week delay ($M = 7.07$, $Med = 6.94$, $SD = 0.36$). We found a significant effect of group ($F(3, 410) = 3.13$, $p < 0.026$, $\eta_p^2 = 0.02$).

As preregistered, we compared the consensus group ($M_{diffT3T1}$ = 4.27, SD = 15.82) to the control group ($M_{diffT3T1}$ = -0.08, SD = 13.61), and found the T3-T1 difference-in-differences test to indicate a significant effect ($t(204)$ = 2.13, $M_{diff\text{-}in\text{-}diffs}$ = 4.35, 95% CI = [0.33, 8.37], p = .034, d = 0.29), which indicates that the positive effect of consensus messaging remains intact.

While not preregistered, after observing a descriptive T3-T2 decrease in the consensus group ($M_{diffT3T2}$ = -4.78, SD = 13.58), we wanted to evaluate whether this could imply that there is a partial decay of the consensus effect. Compared to the control group ($M_{diffT3T2,control}$ = -0.76, SD = 13.50), we found a significant decay ($t(207)$ = -2.15, $M_{diff\text{-}in\text{-}diffs}$ = -4.03, 95% CI = [-7.72, -0.33], p = .033, d = -0.30). More specifically, comparing the consensus-control difference scores of T3-T1 to T2-T1, we found that the consensus messaging effect decays by 48% over the course of one week.[17] For a complete overview of all raw means per group and within-group difference score analyses, see Supplementary Analysis S2.

H3: The Misinformation Effect

The third hypothesis stated there would be a negative effect of exposure to a doubt-sowing misinformation message one week after being exposed to the consensus message, resulting in a complete neutralisation of the initial positive effect. As preregistered, we compared the (false-)balanced group ($M_{diffT3T1}$ = -1.78, SD = 22.83) to the control group ($M_{diffT3T1}$ = -0.08, SD = 13.61), and found no effect ($t(163)$ = -0.65, $M_{diff\text{-}in\text{-}diffs}$ = -1.70, 95% CI [-6.91, 3.50], p = .51, d = -0.09). This indicates that the positive benefit has indeed been eliminated by the misinformation, in line with our hypothesis.

See Figure 4.1.4.3 for a bar chart of within-group differences, which highlights the negative effect of exposure to misinformation in all three groups. See Supplementary Analysis S2 for an overview of all pre, post, and within-group difference scores.

[17] Decay formula used: 1-(Consensus (T3 - T1) - Control (T3 - T1)) / (Consensus (T2 - T1) - Control (T2 - T1)).

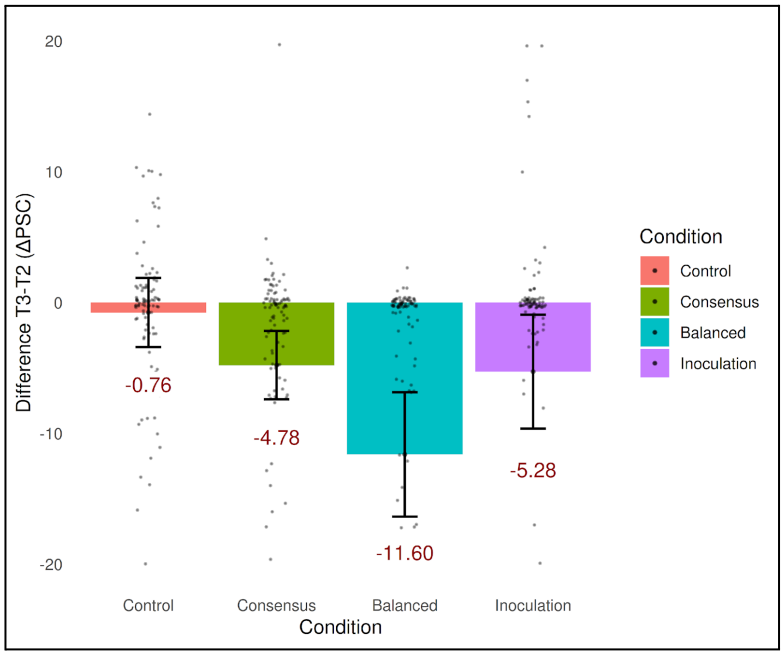

Figure 4.1.4.3. Comparison of T3-T2 *Perceived Scientific Consensus* difference scores (in percentage points) for each group. Error bars represent 95% confidence intervals. $N = 415$.

H4: The Inoculation Effect

Our final hypothesis stated that the inoculation would protect the positive effects of the consensus message from misinformation presented one week later. As preregistered, we did this by comparing the inoculation group ($M_{\text{diffT3T1}} = 4.90$, $SD = 20.60$) to the control group ($M_{\text{diffT3T1}} = -0.08$, $SD = 13.61$), and found the predicted positive effect ($t(104) = 2.06$, $M_{\text{diff-in-diffs}} = 4.97$, 95% CI = [0.20, 9.74], $p = .041$, $d = 0.28$). While not preregistered, as a final robustness check, we also compared the difference-in-differences for the inoculation group ($M_{\text{diffT3T1}} = 4.90$, $SD = 20.60$) compared to the (false-)balanced group ($M_{\text{diffT3T1,}} = -1.78$, $SD = 22.83$), and found a significant effect ($t(200) = 2.20$, $M_{\text{diff-in-diffs}} = 6.68$, 95% CI = [0.70, 12.66], $p = .029$) with an effect size of $d = 0.31$.[18] These results are all in line with our hypothesis and indicate that inoculation is indeed able to protect the positive effects of consensus messaging against later presented misinformation.

[18] Although not preregistered, consistent with Williams and Bond (2020), we found no significant difference between the consensus-only ($M_{\text{diffT3T1}} = 4.27$, $SD = 15.82$) and the inoculation condition ($M_{\text{diffT3T1}} = 4.90$, $SD = 20.60$): $t(195) = 0.25$, $M_{\text{diff-in-diffs}} = 0.62$, 95% CI = [-4.37, 5.61], $p = .81$.

For a measure of the protection percentage of the inoculation effect, we calculated the consensus effect decay rate for the inoculation group and compared this to the decay rate in the consensus-only group.[19] We found that the consensus effect decays at the same rate (48%) for the inoculation group as for the consensus-only group, without being influenced by the misinformation. At the same time, we found a full decay (100%) of the consensus effect in the false-balance group that received misinformation, but no inoculation. This indicates that inoculation was able to eliminate the negative misinformation effect in its entirety. Given that the misinformation is presented one week after the inoculation, this demonstrates 100% protection with a one-week delay between inoculation and attack.

See Figure 4.1.4.4 for a graph visualising the difference scores in the groups over time, and Supplementary Figure S1 for a bar plot of all conditions and test dates combined. A schematic overview of all difference scores can be found in Supplementary Analysis S2.

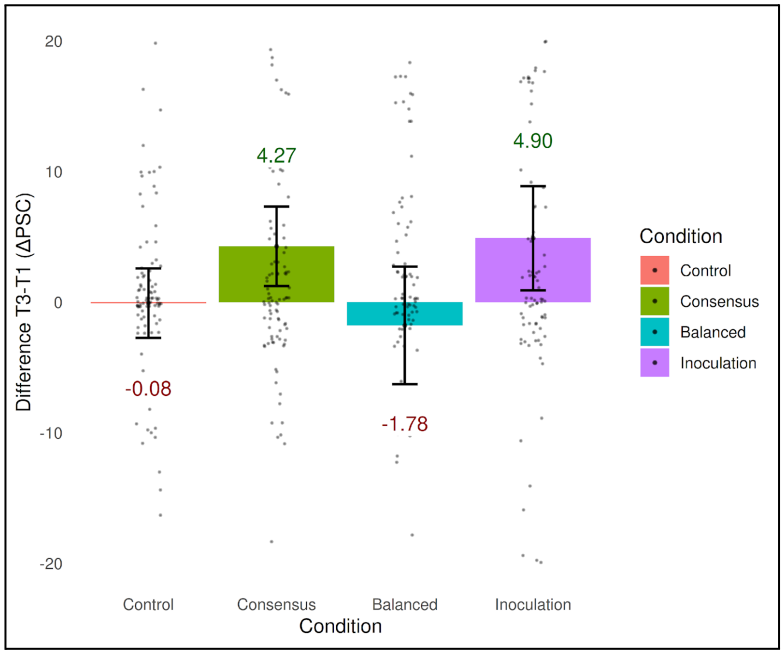

Figure 4.1.4.4. Comparison of T3-T1 *Perceived Scientific Consensus* difference scores (in percentage points) for each group. Error bars represent 95% confidence intervals. $N = 415$.[20]

[19] Retention formula used: [(Inoculation (T3 - T1) - Control (T3 - T1)) / (Inoculation (T2 - T1) - Control (T2 - T1)] / [(Consensus (T3 - T1) - Control (T3 - T1)) / (Consensus (T2 - T1) - Control (T2 - T1)].
[20] A bar plot with all conditions on all different test times can be found in Supplementary Figure S1.

Exploratory

In line with the original study by van der Linden et al. (2017), we examined potential differences by ideology. First, to assess whether the inoculation effect stays intact across the ideological spectrum, we performed a linear regression of the T3-T1 PSC difference scores (within the inoculation group) on ideology, and found no significant effect ($F(1, 103) = 2.63$, $\beta = 0.16$, 95% CI [-2.35, 2.67], $p = .11$, $R^2 = 2\%$). This is in line with the original study.

We found similar stability across party affiliation (Democrat, Independent, Republican), but an uneven starting point ($M_{\text{Democrat}} = 87.96$, $M_{\text{Independent}} = 80.91$, $M_{\text{Republican}} = 72.13$; see Figure 4.1.4.5, panel A, Control). People from all parties benefited from the consensus messaging treatment (see Figure 4.1.4.5, panel A, Consensus) and were negatively affected by the misinformation message (see Figure 4.1.4.5, panel A, Balanced), with the largest changes for Republicans. Finally, when inoculated against misinformation, the net result is positive across party affiliation (see Figure 4.1.4.5, panel B, Inoculation). While most of these interpretations are based on descriptive raw means, they are in line with the original findings by van der Linden et al. (2017).

Next, on the basis of the original study by van der Linden et al. (2017) and the hypotheses formulated by Williams and Bond (2020), we expected 1) the consensus treatment to have a greater positive effect on Republicans than Democrats, 2) the misinformation treatment to have a more negative impact on Republicans than Democrats, and 3) Independents to be the least influenced by the misinformation message.

When collapsing all treatment groups we found that the consensus messaging effect for Republicans ($M_{\text{diffT2T1}} = 15.47$, $SD = 18.90$) was indeed more positive compared to Democrats ($M_{\text{diffT2T1}} = 4.97$, $SD = 10.29$): $t(53) = 3.55$, $M_{\text{diff-in-diffs}} = 10.50$, 95% CI [4.57, 16.43], $p < .001$, $d = 0.69$. This finding is in line with van der Linden et al. (2017), but in contrast to the findings by Williams and Bond (2020), who did not find evidence for this

hypothesis (possibly due to low power). We also found that the misinformation treatment descriptively has a more negative effect on Republicans (M_{diffT3T2} = -21.42, *SD* = 39.59) than Democrats (M_{diffT3T2} = -11.51, *SD* = 24.17), but the effect was not significant ($t(13)$ = -0.83, $M_{\text{diff-in-diffs}}$ = -9.91, 95% CI [-35.66, 15.85], p = .42, d = -0.30). These findings are compatible with both the original study by van der Linden et al. (2017) and the findings by Williams and Bond (2020). Finally, we found that Independents are indeed the least influenced by misinformation (M_{diffT3T2} = -8.63, *SD* = 16.41), which is in line with findings by van der Linden et al. (2017), Cook et al. (2017), and Williams and Bond (2020).[21]

[21] Surprisingly, we also found that the misinformation message exerted a descriptively larger negative influence on moderates (M_{diffT3T2} = -19.61, *SD* = 27.81) than on both conservatives (M_{diffT3T2} = -17.69, *SD* = 34.60) and liberals (M_{diffT3T2} = -7.06, *SD* = 17.92). This seems at odds with the consistent finding that Independents are the least influenced by the misinformation and Republicans the most. However, this finding may simply be due to sampling error or differences between *moderates* and *independents*.

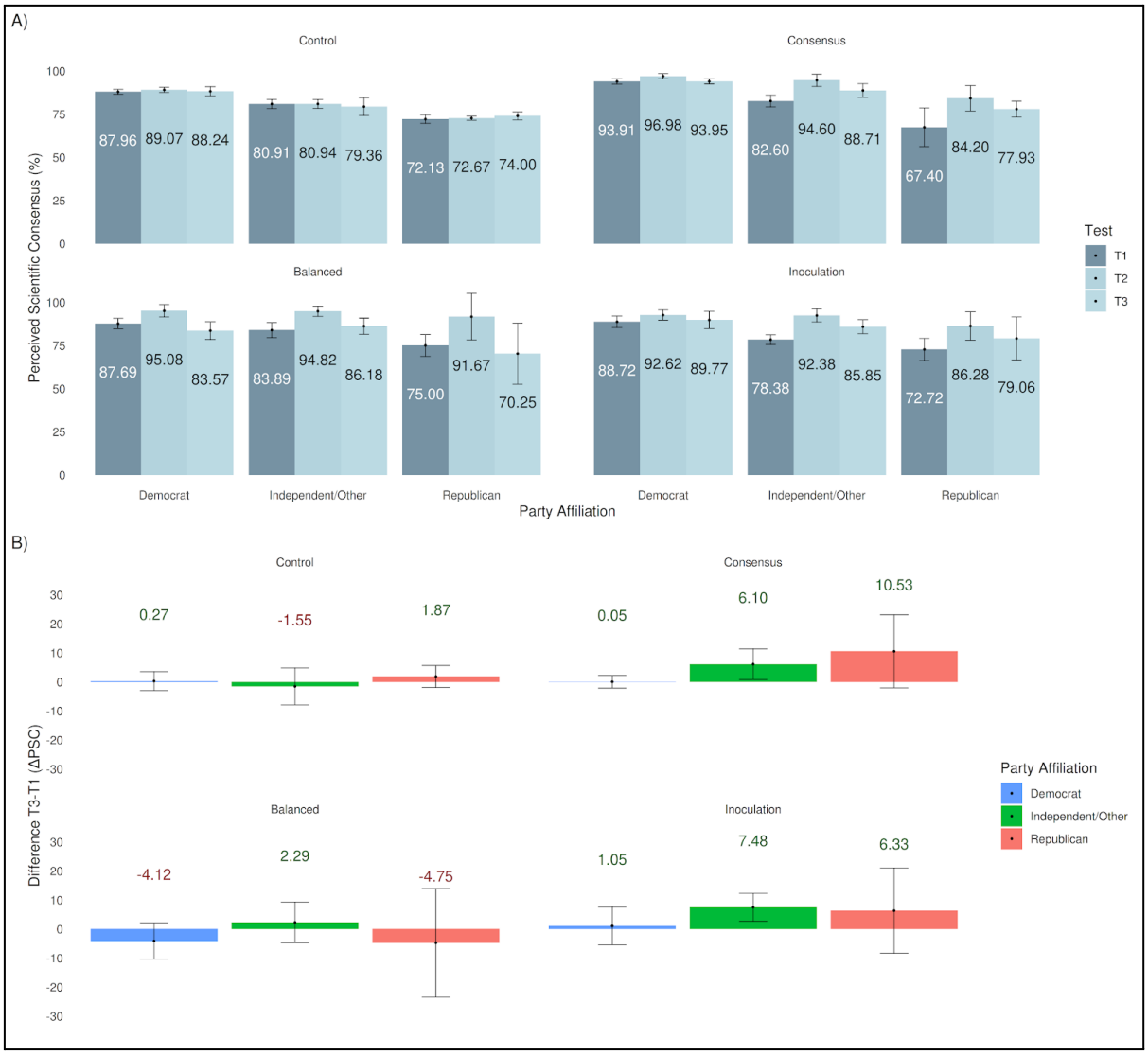

Figure 4.1.4.5. Means for each test (panel A) and T3-T1 difference scores (panel B) of *Perceived Scientific Consensus*, separated by party affiliation ($N = 415$). Error bars represent 95% confidence intervals.

4.1.5 Discussion

In an extended replication study of van der Linden et al. (2017), we conducted a longitudinal experiment where we combined consensus messaging with inoculation and assessed their effects over time. We replicated the initial positive effect of the scientific consensus message by itself and across ideology and party affiliation (Lewandowsky et al., 2013; van der Linden, Leiserowitz, et al., 2015, 2019), but also found that this consensus

effect shows partial decay over the course of one week. We also replicated the finding that the misinformation message counteracts the consensus message and brings perceived scientific consensus back to baseline. Finally, we found that an inoculation message is able to protect the positive effects of the consensus message against doubt-sowing misinformation presented with a one-week delay, without any decay in the inoculation effect over time.

It is important to highlight that a crucial difference between the original study of van der Linden et al. (2017) and the current study is that we introduced a 1-week delay between inoculation and the misinformation attack, assuming that the inoculation would stay intact without reinforcement. Although the initial consensus messaging effect was not as high as in the original study ($d = 0.71$, $d_{\text{vanderLinden}} = 1.23$), and this effect further decayed to an effect of $d = 0.29$ at T3, we still found significant effects for both a consensus effect and an inoculation effect at T3.

Compatible with our findings, an independent replication study of van der Linden et al. (2017) by Williams and Bond (2020) found evidence for both the positive and the protective effects of the consensus message. However, while they found the positive effects of the consensus message to stay significant even after the exposure to the misinformation message, we found a full reversion back to baseline. These findings may indicate that participants in our study gave equal weight to the misinformation as to the facts, which resulted in a net change of zero, consistent with van der Linden et al. (2017). Although Williams and Bond (2020) found a significant inoculation effect compared to a no-treatment control, they found no additional benefit compared to the false-balance condition. These results suggests that the consensus message may have been strong enough to remain significant on its own, and did not need an inoculation as extra protection. We also found a protective benefit of the consensus message on its own, but this effect was less strong, or, alternatively, our misinformation message was more potent. Because in our study no

misinformation-only group was present, this cannot directly be tested in the same way. Nevertheless, our pilot study indicated that the misinformation message caused a significant negative effect (with PSC dropping below baseline level) while in the false balance group we found no change in PSC compared to baseline. This indicates, consistent with Williams and Bond (2020) and Cook et al. (2017), that communicating the consensus on its own may have a positive protective effect against misinformation, even without inoculation (see Supplementary Figure S2). It could be argued that, as the consensus message is presented *before* the misinformation message, the consensus message in itself may have an inoculating feature. Yet, compared to the false-balance group, we found a significant positive effect for the inoculation group ($d_{balanced}$ = -0.09, $d_{inoculation}$ = 0.28), indicating an additional inoculation benefit that is not found in the study by Williams and Bond (2020). However, although the effect of the inoculation group compared to the control group in our study was significant, it was not of the same size as the original study (d = 0.28, $d_{vanderLinden}$ = 0.75), which could partly be explained by the decay in the consensus effect. These results may be consistent with the findings from Niederdeppe et al. (2015), who found that narratives decay faster than inoculation messages.

The different findings and conclusions between our study and the replication by Williams and Bond (2020) pose important questions. One potential explanation is that in their study they found a weaker misinformation effect (d = -0.47, vs $d_{WilliamsBond}$ = -0.25), while we found the same effect as in the original study ($d_{vanderLinden}$ = -0.48). The consensus effect was the same between our studies (d = 0.71, $d_{WilliamsBond}$ = 0.70), in both studies lower than in the original study ($d_{vanderLinden}$ = 1.23). Finally, the biggest differences were found in the false-balance group compared to the consensus group (d = -0.31, $d_{WilliamsBond}$ = 0.09) and the balanced group compared to the inoculation group (d = -0.31, $d_{WilliamsBond}$ = 0.07). We could ask whether a *Prolific* sample is fundamentally different from an *MTurk* sample but looking

at the sample composition this does not seem likely (see Supplementary Information S5). Another explanation is that the specific participants in Williams and Bond (2020) may have been exposed to the treatment and/or misinformation message before. Ceiling effects and timing of the experiments are unlikely to explain the differences as dates and pretest means are similar between our studies. An alternative explanation is a ceiling effect combined with a difference in design. For example, in both studies, a higher pretest score was found for all conditions compared to the original, which had a much larger and more diverse sample (van der Linden et al., 2017). In addition, in our study we present the misinformation message in isolation one week after the consensus message, giving the opportunity for the consensus messaging effect to decay over time and thereby allowing for higher saliency of the misinformation message. We therefore conclude that the most likely explanation is the difference in the misinformation message, which may not have been strong enough in the study by Williams and Bond (2020). As Williams and Bond (2020) note, they also used a slightly different misinformation message, one without a description of the Oregon Petition. For example, in the study by van der Linden et al. (2017), the current study, and Experiment 2 by Cook et al. (2017), a descriptive text was presented together with the misinformation message (see Supplementary Information S2). In all three of these studies, similar misinformation and inoculation effects were found. In contrast, in Cook et al. (2017), Experiment 1, a lower misinformation effect was reported, and no additional inoculation benefit was found in comparison to the consensus-only condition, which aligns more closely with the results by Williams and Bond (2020). We therefore recommend future inoculation studies to carefully pre-test and evaluate the efficacy of the misinformation stimuli.

In both our study and the replication by Williams and Bond (2020) the consensus and inoculation effects were not affected by political ideology. As our experiments focus on the heavily polarized topic of climate change, our studies are in line with growing evidence that

inoculation theory is applicable in the context of contested issues and not highly vulnerable to reactance or backfire effects (Cook et al., 2017; van der Linden et al., 2017; Williams & Bond, 2020). Indeed, jointly, these findings are consistent with other recent work which does not find a backfire effect (Guess & Coppock, 2018; T. Wood & Porter, 2019), in contrast to identity protection theories where backfire effects would be expected in the form of attitude polarization (Hart & Nisbet, 2012; Kahan et al., 2011; Nyhan & Reifler, 2010).

Our study does not come without limitations. For instance, we limited our study to a US-only convenience sample from Prolific. Future studies need to investigate whether these results replicate in representative samples and in non-WEIRD countries and regions with different cultural worldviews (Henrich et al., 2010; Tam & Milfont, in press). In our power calculations we did not take into account the high level of decay (48%) found in the consensus messaging effect. Future longitudinal studies will need to account for the smaller longitudinal effects of $d = 0.28$ found in this study.

Future research is needed to map the decay process of the consensus and inoculation effects more meticulously using longitudinal studies with a higher sample size and more timepoints. One could argue that a one-week retention effect is not sufficient to talk about *long-term* effectiveness in practical terms. It would be useful to know whether the effects last for more than one month, as this would both provide evidence for true long-term memory consolidation (Frankland & Bontempi, 2005) and have practical consequences for policy implementations (i.e., minimal resources needed for long-lasting effects). If decay is found, it would be valuable to investigate the decay function and gain insight into how decay could be prevented. One potential avenue to explore is to combine the consensus message with inoculation "booster" sessions (Compton & Pfau, 2005; McGuire, 1961). Just as in the biomedical metaphor of a vaccine, it may be necessary to give the cognitive immune system a regular top-up to remind it what to protect against (Ivanov & Parrott, 2017). Potential

"booster shots" can be either a repetition of the inoculation message, the consensus message, or the misinformation message, or a combination of these (Ivanov et al., 2009; Pfau et al., 2005; Pryor & Steinfatt, 1978).

Overall, these results provide support for the implementation of inoculation-inspired interventions. For example, one approach to counter climate misinformation in the real-world is to create game-based inoculation interventions that can be used directly by climate change communicators in educational, professional, and policy settings. Examples include the popular fake news game, *Bad News* (Basol et al., 2020; Roozenbeek & van der Linden, 2019a) and the smartphone application *Climate Change vs. Cranky Uncle* (Goering, 2019).

Conclusion

Although not all conclusions in our study are straightforward, we have evidence to conclude that both consensus messaging and inoculation theory are effective methods to combat climate change misinformation, and that the longevity of inoculation spans for at least one week. Protecting newly changed beliefs on polarized topics without decay extends the initial predictions of inoculation theory, while building upon the same foundations. This allows for the development of new strategies to combat climate change misinformation, with at least some resistance over time. As this intervention has shown promise in three independent studies at different time-points and across different ideological groups, we invite policy makers and communicators to start evaluating inoculation interventions in the field. The most prevalent forms of misinformation could be identified and severely weakened doses could be tested and distilled into inoculation messages disseminated via social media platforms, news articles, and press conferences. As for most large scientific theories, more investigation and field studies are needed to establish the boundary conditions of long-term resistance, and its practical utility for public policy will have to be evaluated through evidence-based policy applications.

Study 2

4.2 The Memory-Motivation Model of Inoculation: Message-Based Interventions[22]

4.2.1 Abstract

The effectiveness of message-based inoculation and consensus interventions has been shown in various previous studies, including Study 1 of this dissertation. In Study 2, we employed a large longitudinal experiment (N = 1,825) to go one step further to explore the underlying mechanisms of the inoculation effect, and in particular, its long-term effectiveness, as well as the effectiveness of "booster" interventions. We show that message-based inoculation interventions can offer protection against misinformation for at least 29 days without any booster intervention, but that there is a decay in the effect over time that is best explained by forgetting, and can be remedied by a short memory strengthening booster intervention. Finally, we tested the memory-motivation model, and found evidence for the role of both memory and motivation in inoculation effects, with memory being the most dominant factor.

4.2.2 Introduction

In the current study we built further on the insights from Study 1, using the same climate change (CC) inoculation design but with a longitudinal experiment that features a wider range of time points and measures. This study allowed us to 1) draw a more detailed decay curve, 2) answer the core research questions about the underlying mechanisms in a specific, passive, therapeutic inoculation design, explore the role of threat, motivation, and memory as a predictor of inoculation performance and longevity, as well as test the validity

[22] Study 2, as depicted here, is a manuscript in preparation as *"The Long-Term Effectiveness of Consensus-Based Inoculation Messages: Mechanisms"*. The paper was written in collaboration with Professor Jon Simons (University of Cambridge), Dr Jon Roozenbeek (University of Cambridge), and Professor Sander van der Linden (University of Cambridge). I am the sole first author on this work.

of the memory-motivation model of inoculation, and 3) test the effectiveness of "booster treatments" and explore the memory strengthening hypothesis.

We hypothesized that we would replicate the negative effect of misinformation on the perceived scientific consensus on human-caused global warming (PSC; H1), as well as the protection effect that an inoculation intervention confers (H2). In the new longitudinal hypotheses we chose for a follow-up after 10 days (T2) and after 30 days (T3). We chose these time points as we know from the previous study that there is no significant decay after one week, but the literature suggests that typically decay can be detected when looking at 2 weeks after the intervention or beyond (Banas & Rains, 2010; Zerback et al., 2021). This allows us to test the limits of the effect with the hypothesis that the effect may still be intact after 10 days (H3), but not after 30 days (T3; H4). Meanwhile, it allows us to test the effectiveness of a booster intervention in the form of a repetition of the original intervention at T2, which we expect to top up the effect and reduce its decay at T3, which is 30 days after T1 or 20 days after the booster at T2 (H5). Finally, we provided three hypotheses to test the memory-motivation model of inoculation, with the booster expected to improve memory (H6) and motivation (H7), and the inoculation effect to be mediated by memory and motivation (H8). The primary purpose was to gain insights into the memory-motivation theory, but also to fill the gap in the literature on booster shots, as *"much more needs to be learned about the best way to structure and time booster messages"* (Ivanov & Parrott, 2017, p. 23). The list of the preregistered hypotheses can be found in Table 4.2.2.1.

The experiment was preregistered on AsPredicted at https://aspredicted.org/GPR_5FB. All materials, survey files, analysis scripts, and raw and clean datasets are available on the OSF repository for this study at https://osf.io/9zxje/?view_only=44a8556694b54d09a2e2a9875071de2f.

Table 4.2.2.1
Hypotheses of Study 2

#	Hypothesis
H1	Exposure to misinformation about climate change in the form of a false petition decreases the perceived scientific consensus (PSC) on global warming.
H2	Inoculated individuals do not negatively change their perceived scientific consensus (PSC) on global warming after exposure to a misinformation message in the form of a false petition.
H3	The inoculation effect described in H2 remains significant for at least 10 days (T2).
H4	The inoculation effect described in H2 is no longer significant after 30 days (T3).
H5	The inoculation effect described in H2 is still significant after 30 days (T3), when individuals are exposed to a second inoculation message after 10 days (T2).
H6	Groups exposed to a second inoculation message after 10 days (T2) show increased memory of the inoculation intervention after 30 days (T3) compared to those exposed to only one inoculation message.
H7	Groups exposed to a second inoculation message after 10 days (T2) show increased motivational threat after 30 days (T3) compared to those exposed to only one inoculation message.
H8	The inoculation effect [H8a] immediately after intervention (T1), [H8b] after 10 days (T2), and [H8c] after 30 days (T3) is influenced directly by memory and motivation, and indirectly by the inoculation intervention (mediated by memory and motivation).

4.2.3 Methods

Design, Sample, and Procedure

The study followed the same procedures as Study 1, including the same misinformation, consensus, and inoculation messages, but with more and longer time periods (T2 = 10 days, T3 = 30 days), a set of new measures, and a condition that includes a booster treatment. In addition, for this study, the consensus and inoculation messages were no longer separated but combined on a single page, and represent the "inoculation" group. We recruited a high-powered sample (power = 95%, α = 5%, decay = 40%, attrition = 30%) of US participants aged 18 or older through Prolific (N = 2,657). As preregistered, participants were excluded when they 1) failed both attention checks (e.g., *"The colour test is simple. When asked for your favourite colour you must not select the word puce, but you have to select the word blue. Based on the text you read above, what colour have you been asked to select?"*, [multiple choice, 5 colors]), 2) participated in the survey multiple times, or 3) did not

complete the entire survey.[23] We also excluded participants who did not participate within a window of 3 days from the intended participation date (i.e., 3 days before or after).[24] This led to a final sample size of $N = 1{,}825$, with an average of 260 participants per group, slightly below the intended $n = 328$ due to a higher-than-expected attrition rate ($T2_{Attrition} = 31\%$, $T3_{Attrition} = 52\%$). Of the final sample, 49.21% identified as male (48.22% as female; 2.03% as non-binary; 0.33% as transgender, 0.22% as "other"), the average age was 35.79 ($SD = 13.07$, $Mdn = 33$), 58.69% had a higher education degree (or higher), 62.58% identified as left-wing (22.47% as centrist; 14.96% as right-wing), 48.99% identified most as Democrat (29.48% as Independent; 10.47% as Republican), 65.59% used social media multiple times a day (19.29% once a day, 7.29% weekly, 4.99% less often than weekly, 2.85% never), and 22.19% used Twitter multiple times a day (13.86% once a day, 12.06% weekly, 19.07% less often than weekly, 32.82% never). The participants were randomly allocated to one of three interventions: a word sorting task (Control), the inoculation message (Inoc), or the inoculation with a booster inoculation at 10 days (Inoc-B). We also separated each time point by recruiting a separate sample for each condition, to avoid effects of repeated testing (i.e., each participant only ever received one post-test, at one time point depending on the group they were allocated), leading to a total of 7 groups. The booster treatment employed in this study is an exact repetition of the original intervention (i.e., rehearsal). All participants received the Oregon Petition misinformation message just before the posttest (cf. Study 1). When not otherwise specified, when we refer to "T1", we refer to the posttest at T1. For a complete overview of the study design, see Table 4.2.3.1.

[23] We also preregistered that we would exclude participants who failed the manipulation check. However, the final version of the study did not contain a manipulation check so we were unable to do this.

[24] We preregistered that we would exclude participants who did not participate in the follow-up within 5 days after the invitation. However, as the invitations were manual and grouped together, we invited participants 1–3 days earlier than the intended follow-up time. Therefore we have changed the exclusion window to 3 days before or after the intended follow-up time instead.

Table 4.2.3.1
Experimental Setup of Climate Change Decay Study 2

Condition	Time		
	Baseline (T1)	**10 Days (T2)**	**30 Days (T3)**
Control-T1 (*n* = 302)	PSC | **C** | **M** | PSC	-	-
Control-T2 (*n* = 263)	PSC | **C**	**M** | PSC	-
Control-T3 (*n* = 206)	PSC | **C**	-	**M** | PSC
Inoc-T1 (*n* = 317)	PSC | **I** | **M** | PSC	-	-
Inoc-T2 (*n* = 254)	PSC | **I**	**M** | PSC	-
Inoc-T3 (*n* = 239)	PSC | **I**	-	**M** | PSC
Inoc-B-T3 (*n* = 244)	PSC | **I**	**I**	**M** | PSC

Note. PSC = measure of perceived scientific consensus. C = control task (word sorting). M = misinformation treatment. I = inoculation treatment.

Materials and Measures

The main dependent variable for this study is the perceived scientific consensus on human-caused global warming, presented on a percentage slider scale (M = 84.10, SD = 16.77). The question asked to the participants is *"To the best of your knowledge, what percentage of climate scientists have concluded that human-caused climate change is happening? (0% to 100%)"*.

This study also introduced a new set of memory and motivation variables, as well as a range of measures that are related to inoculation effects and the memory-motivation model. Our main measure for memory was an objective, performance-based, inoculation intervention content recall test that we designed for this study. It included 12 objective questions, including 8 yes-or-no questions (e.g., *Which of the following did you learn about in the messages from the first part of the survey?*; *False petitions*, *Yes/No*) and 4 multiple choice questions (e.g., *What was the message from the first part of the survey about?*; a – *The scientific consensus on climate change*, b – *Financial policy in the United States*, c – *The*

political side of bowling, d – *Vaccination intentions*, e – *None of these options is correct*), which were combined into an index variable that we refer to as "memory" in this study (0–12; $M = 7.54$, $SD = 2.09$). For exploratory purposes, a set of subjective memory measures specifically created for this study was included as well, including *self-reported remembrance* (e.g., *"How well do you remember the messages about climate change you saw earlier in the survey?"*, Likert scale 1–7; $M = 3.96$, $SD = 1.86$), 4 open questions (e.g., *"What do you remember about the first half of the survey?"*), and 3 questions related to interference that were combined into an interference index (e.g., *"In the past two weeks, I have heard conflicting arguments about climate change"*; Likert scale 1–7, *Not at all true–Very true*; $M = 8.93$, $SD = 4.66$).

Next to memory questions, we implemented a range of motivation measures. Our main measure for motivation was *motivational threat* (adapted version of measures used by Banas & Richards, 2017; Richards & Banas, 2018), which is seen as the most predictive measure of threat-based motivation for inoculation-induced resistance to misinformation (Banas & Richards, 2017; Richards & Banas, 2018). We calculated this variable using a mean index of 3 Likert scale questions (e.g., *"Thinking about climate change misinformation motivates me to resist misinformation"*, 1–7, *Strongly disagree–Strongly agree*; $M = 5.20$, $SD = 1.50$). This is what we refer to as "motivation" for the rest of this chapter. In addition, as exploratory measures for the memory-motivation model, we also included measures for *apprehensive threat*, *fear*, *issue involvement*, *issue accessibility*, and *issue talk* (adapted versions based on the measures used by Banas & Richards, 2017; Dillingham & Ivanov, 2016; Ivanov et al., 2012; Pfau et al., 2003, 2004, 2005; Richards & Banas, 2018). We measured *apprehensive threat* using a mean index of 6 Likert-scale questions (e.g., *"Thinking about climate change misinformation I feel threatened"*; 1–7, *Strongly disagree–Strongly agree*; $M = 3.92$, $SD = 1.73$), *fear* using a mean index of 3 Likert scale questions (e.g.,

"Thinking about climate change misinformation I feel fearful"; 1–7, *None of this feeling–A great deal of this feeling*; M = 3.96, SD = 1.92), *issue involvement* using an index score of "choose one option from this pair" questions (e.g., *"Which option of each pair best describes how much climate change means to you?"*; *Insignificant*, *Significant*; converted to 1–7 score, M = 6.27, SD = 1.76), *issue accessibility* using a single Likert scale item (*"Compared to other issues, how often do you think about climate change?*"; 1–7, *Never–Very often*; M = 3.95, SD = 1.60), and *issue talk* using an index of 3 questions converted to a score from 1–7 (M = 2.23, SD = 1.23), including 2 Likert scale questions (e.g., *"In the past two weeks, how often did you talk about or discuss climate change with other people"*; 1–7, *Never–Very often*) and 2 choice option list questions (e.g., *"In the past two weeks, how many times did you talk about or discuss climate change?*"; *0*, *1*, *2*, *3*, *4*, *5*, *More than 5*).

The original survey files as well as a printout of the full survey can be found on the OSF repository for this study at https://osf.io/9zxje/?view_only=44a8556694b54d09a2e2a9875071de2f.

4.2.4 Results

Hypothesis Tests

Main Effect

As preregistered, we started by testing the main effect of the misinformation message [H1] and the main effect of the inoculation message [H2] using a paired t-test (pre vs. post). We found that, in line with the hypotheses, the misinformation message had a negative effect on the perceived scientific consensus (PSC), [H1] M_{diff} = -4.80, 95% CI [-7.10, -2.50], $t(301)$ = -4.11, $p < .001$, d = -0.237, 95% CI [-0.351, -0.122], while when an inoculation message was shown before the misinformation message, there was no negative effect, or better, there was a positive effect, [H2] M_{diff} = 7.34, 95% CI [5.59, 9.10], $t(316) = 8.24$, $p < .001$, d = 0.463, 95% CI [0.347, 0.579].

Decay Analysis

After replicating the main effects, we investigated the effect retention at T2 (*Mdn* = 8 days) and T3 (*Mdn* = 29 days) using an ANCOVA with pretest PSC as a covariate, posttest PSC as the outcome variable, and intervention as an independent variable.[25] We first found that the inoculation effect was still significant at 8 days, [H3] $F(1, 514) = 18.94$, $p < .001$, $d = 0.384$, 95% CI [0.209, 0.558], replicating the findings from Study 2. For the analyses at 29 days, we first confirmed a significant omnibus test for the intervention variable, $F(2, 685) = 12.63$, $p < .001$, and then found that the effect at 29 days was still significant with a smaller effect, [H4] $t(685) = 2.96$, $p_{tukey} = .009$, $d = 0.281$, 95% CI [0.094, 0.468]. As we expected the inoculation effect after 29 days to no longer be significant, this result provides evidence against H4. See Figure 4.2.4.1 (Panel A) for a visual plot of the inoculation effects over time.

[25] The analyses used for H3–H7 are slightly different from the preregistered analyses. The preregistration mentioned a repeated measures ANCOVA but as we do not have fully balanced conditions and we have separate groups for each time point we instead use a separate ANCOVA for each time point.

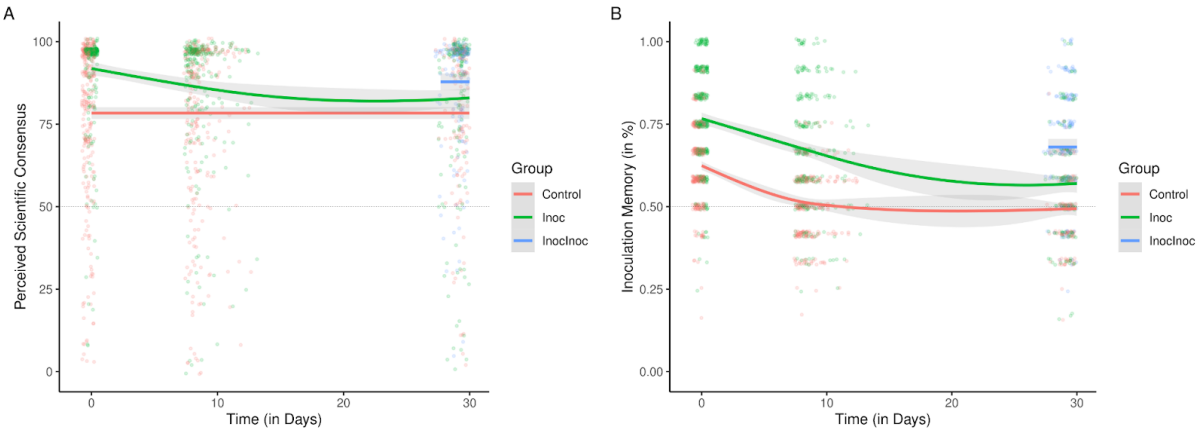

Figure 4.2.4.1. Visual plot of the perceived scientific consensus after exposure to a misinformation message (Panel A) and objective memory of the inoculation intervention (Panel B) over time. The *InocInoc* group depicts the booster condition that received the inoculation message twice. Error bands represent the standard error. $N = 1{,}825$.

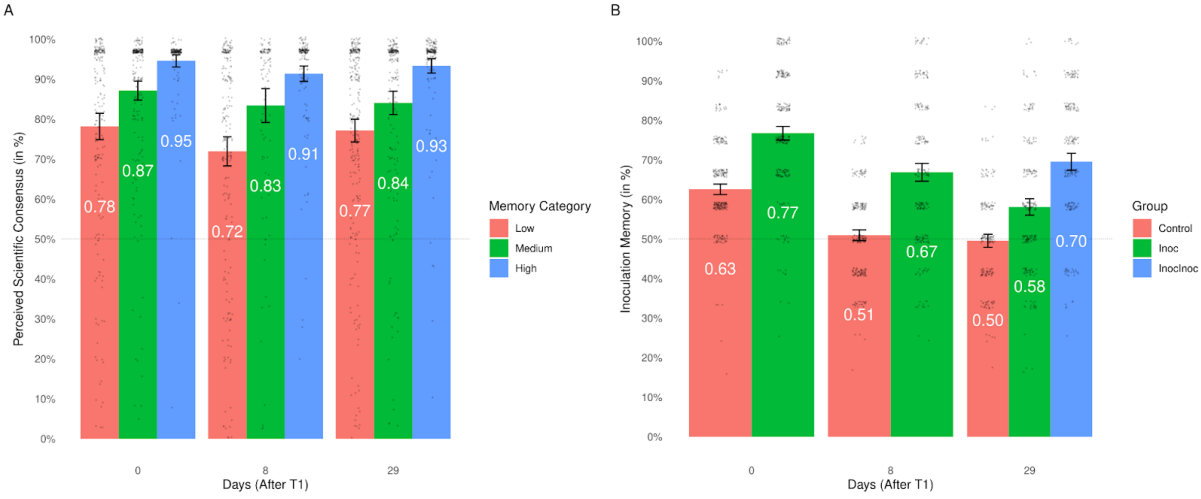

Figure 4.2.4.2. Panel A represents the perceived scientific consensus for those scoring low, medium, or high on memory in the inoculation group, split by date ($N = 1{,}054$). Panel B represents the objective memory of the inoculation intervention in each group, with a visible memory boost for the group that received a second inoculation after 8 days ($N = 1{,}852$). The *InocInoc* group depicts the booster condition that received the inoculation message twice. Error bars represent the 95% Confidence Interval.

Booster Analysis

To test H5, we look at the inoculation effect at T3 (29 days) for participants who took part in the booster intervention as compared to the control group, and found a significant medium effect, [H5] $t(685) = 5.02$, $p_{tukey} < .001$, $d = 0.475$, 95% CI [0.287, 0.662], in line with the hypothesis. Although not preregistered, we also looked at the contrast between the booster group and the inoculation group, and found no significant effect, $t(685) = 2.13$, $p_{tukey} = .085$, $d = 0.194$, 95% CI [0.015, 0.373]. See Figure 4.2.4.2 for a plot of the effect of memory on the perceived scientific consensus, and the memory boost provided by a second inoculation after 8 days.[26]

Memory-Motivation Model

Using an ANOVA analysis with intervention as the independent variable and objective memory of the inoculation intervention as the outcome variable for H6, and a separate ANOVA with motivation as the outcome variable for H7, we tested the direct effects of the booster condition on the two mediators in the memory-motivation model at T3 (29 days). For this analysis we only looked at T3 as participants in the booster condition only received the posttest questions at T3 (at T1 and at T2 they received the inoculation and booster interventions without posttest measurement). We first found that the omnibus test for the intervention was significant for memory, $F(2, 686) = 95.28$, $p < .001$, in line with H6, but not for motivation, $F(2, 686) = 0.31$, $p = 0.731$, leading to the rejection of H7. The contrast between the double inoculation (booster) group and the single inoculation group for objective memory showed a strong significant effect of the booster intervention, [H6] $t(686) = 8.15$, $p_{tukey} < .001$, $d = 0.741$, 95% CI [0.558, 0.924], in line with H6.

As preregistered, we also tested an approximation of the memory-motivation model using an SEM analysis with the *lavaan* package in R (Rosseel, 2012). In this model we

[26] The control group on the memory graph (Panel B) depicts guessing. One potential reason for the slight inflation at T1 is that also in the control group participants were asked questions about the scientific consensus on human-caused global warming, and this may have influenced guessing performance.

entered inoculation at T1 (yes/no) as a predictor variable, motivational threat and objective memory as mediator variables, and PSC as an outcome variable. The purpose of this model is to disentangle the direct and indirect effects, and not to compare with other models. As all variables were entered as observed variables (not as latent variables) and we allowed all variables in the model to be related to each other, the model is saturated and thus has perfect fit values. We created a separate model for each time point as for each time point we had a different sample and thus we could not calculate the paths between the variables at different time points. See Figure 4.2.4.3 for a schematic depiction of the T3 model and Table 4.2.4.4 for its estimates.

We found that, in line with the hypothesis, that there was a direct effect of inoculation memory on the PSC at T1, $z = 5.51$, $p < .001$, $\beta = 0.230$, 95% CI [0.148, 0.311], at T2 (8 days), $z = 7.93$, $p < .001$, $\beta = 0.316$, 95% CI [0.227, 0.406], and at T3 (29 days), $z = 7.86$, $p < .001$, $\beta = 0.291$, 95% CI [0.218, 0.363]. Similarly, a direct effect was found of motivation on the PSC at T1, $z = 4.77$, $p < .001$, $\beta = 0.177$, 95% CI [0.104, 0.250], T2, $z = 4.96$, $p < .001$, $\beta = 0.200$, 95% CI [0.121, 0.280], and at T3, $z = 2.51$, $p = .012$, $\beta = 0.088$, 95% CI [0.019, 0.157]. Meanwhile, the inoculation intervention had a direct influence on memory at T1, $z = 12.93$, $p < .001$, $\beta = 0.907$, 95% CI [0.769, 1.044], at T2, $z = 11.97$, $p < .001$, $\beta = 0.926$, 95% CI [0.774, 1.077], and at T3, $z = 10.63$, $p < .001$, $\beta = 0.816$, 95% CI [0.665, 0.966]. Also motivation had an impact on memory at T1, $z = 4.94$, $p < .001$, $\beta = 0.173$, 95% CI [0.105, 0.242], at T2, $z = 2.45$, $p = .014$, $\beta = 0.095$, 95% CI [0.019, 0.170], and at T3, $z = 2.51$, $p = .012$, $\beta = 0.088$, 95% CI [0.019, 0.157]. The intervention did not have a direct influence on motivation at T1, $z = 0.51$, $p = .608$, $\beta = 0.041$, 95% CI [-0.116, 0.199], at T2, $z = 1.26$, $p = .207$, $\beta = 0.111$, 95% CI [-0.061, 0.283], or at T3, $z = 0.78$, $p = 0.434$, $\beta = 0.065$, 95% CI [-0.098, 0.228]. Finally, the inoculation intervention had an indirect influence on the PSC mediated by memory at T1, [H8a] $z = 5.07$, $p < .001$, $\beta = 0.208$, 95% CI [0.128, 0.289], T2,

[H8b] $z = 6.00$, $p < .001$, $\beta = 0.293$, 95% CI [0.197, 0.388], and at T3, [H8c] $z = 6.32$, $p < .001$, $\beta = 0.237$, 95% CI [0.164, 0.311], providing evidence in line with the memory-motivation model.

While not preregistered, to investigate the nature of the mediation model further, we also looked at the direct effect of inoculation on the PSC at T3, and found that it was not significant $z = 0.95$, $p = .341$, $\beta = 0.076$, 95% CI [-0.082, 0.233], while the indirect effect was significant $z = 4.06$, $p < .001$, $\beta = 0.334$, 95% CI [0.173, 0.495]. This provides evidence for full mediation.

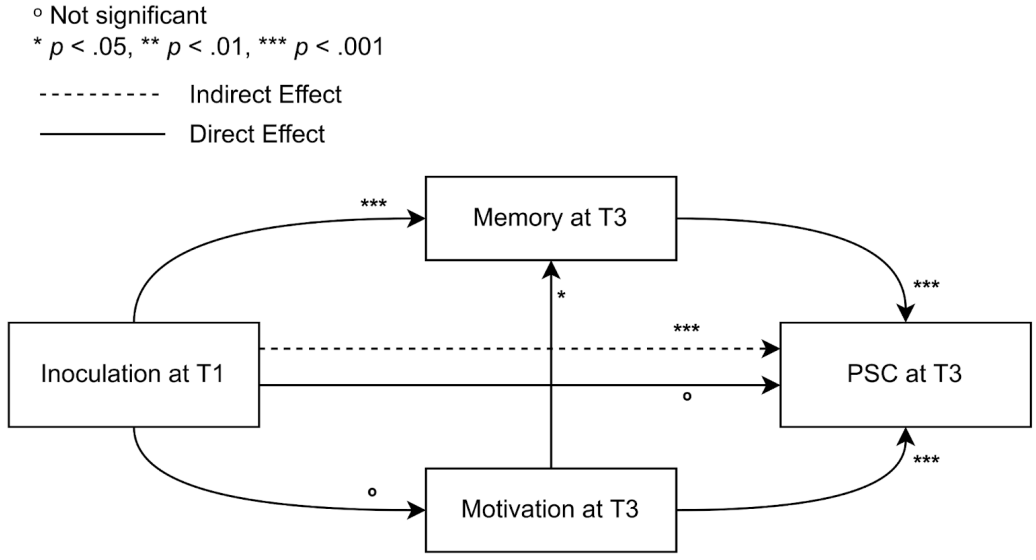

Figure 4.2.4.3. SEM Analysis of the memory-motivation model at T3 in Study 2.

Table 4.2.4.4
Memory-Motivation Model Estimates at T3 in Study 2, N = *689*

Effect	*z*	*p*	β	95% CI		*SE*
				LL	*UL*	
Indirect						
Inoc.T1 ⇒ Memory.T3 ⇒ PSC.T3	**6.323**	**< .001**	**0.237**	**0.164**	**0.311**	**0.038**
Inoc.T1 ⇒ Motivation.T3 ⇒ PSC.T3	0.779	.436	0.019	-0.029	0.068	0.025
Inoc.T1 ⇒ Motivation.T3 ⇒ Memory.T3 ⇒ PSC.T3	0.744	.457	0.002	-0.003	0.006	0.002
Component						
Inoc.T1 ⇒ Memory.T3	**10.634**	**< .001**	**0.816**	**0.665**	**0.966**	**0.077**
Memory.T3 ⇒ PSC.T3	**7.864**	**< .001**	**0.291**	**0.218**	**0.363**	**0.037**
Inoc.T1 ⇒ Motivation.T3	0.782	.434	0.065	-0.098	0.228	0.083
Motivation.T3 ⇒ PSC.T3	**8.632**	**< .001**	**0.296**	**0.229**	**0.363**	**0.034**
Motivation.T3 ⇒ Memory.T3	**2.509**	**.012**	**0.088**	**0.019**	**0.157**	**0.035**
Direct						
Inoc.T1 ⇒ PSC.T3	0.945	.345	0.076	-0.082	0.233	0.080
Total						
Inoc.T1 ⇒ PSC.T3	**4.063**	**< .001**	**0.334**	**0.173**	**0.495**	**0.082**

Exploratory Analyses

Dominance Analysis

Looking further into the mechanisms of the inoculation effect, setting out to find out what the strongest predictor is of the inoculation effect, we implement a dominance analysis with the T3 data of a wide range of predictors of the inoculation outcome mentioned in the literature. Dominance analysis is a method to investigate the relative importance of each predictor variable in a regression model by calculating the additional variance explained (R^2) of each variable in all possible model combinations with these variables and then performing pairwise comparisons for each of these subsets to establish which variable was more important (i.e., more dominant), leading to a percentage of the cases where one variable was dominant above the other variables (Budescu, 1993). This allowed us to identify which predictors were the most essential predictors. We use the T3 data as this time point is most relevant in terms of uncovering the mechanisms behind the long-term effectiveness. The analysis demonstrated that memory was by far the most dominant predictor of the inoculation

effect (82%). See Table 4.2.4.5 for an overview. See Figure 4.2.4.1 (Panel A & B) for a comparison of the decay cure of the PSC and the forgetting curve of the inoculation memory. See Figure 4.2.4.6 for a plot of the correlation between memory (Panel A) and motivation (Panel B) and the PSC.

Table 4.2.4.5

Dominance Analysis in Study 2, at T3, N = *689*

Variable	Dominance
Inoculation Memory	82%
Issue Involvement	4%
Issue Accessibility	4%
Apprehensive Threat	4%
Motivational Threat	2%
Self-Reported Remembrance	1%
Fear	1%
Issue Talk	1%

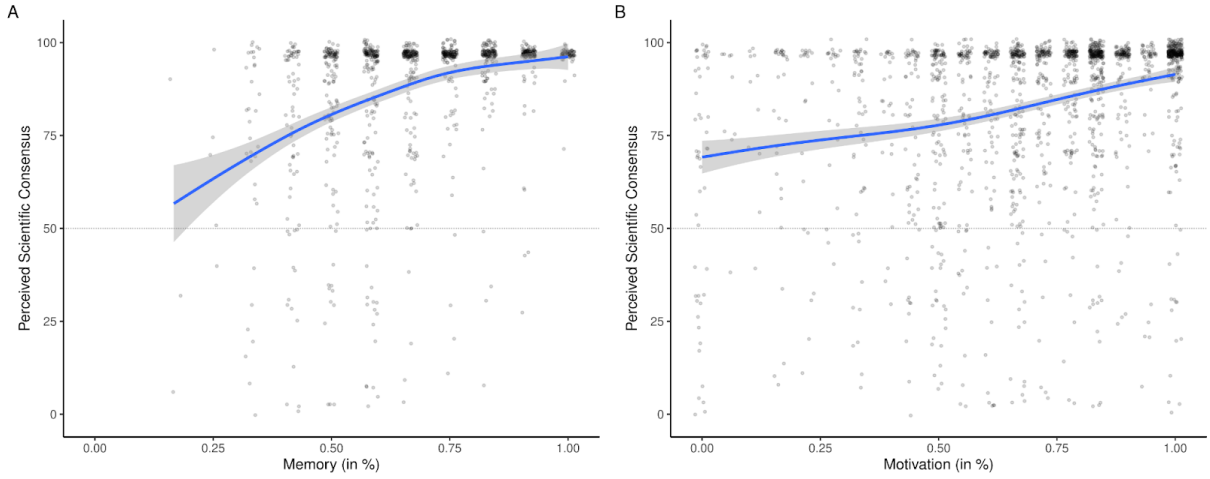

Figure 4.2.4.6. Correlation plot of the perceived scientific consensus after exposure to a misinformation message with inoculation memory (Panel A) and motivation (Panel B). Error bands represent the standard error. $N = 1{,}825$.

4.2.5 Discussion

With Study 2 we investigated the longevity of the inoculation effect in a specific, passive, therapeutic, text-based inoculation intervention, and investigated its underlying mechanisms. We first found that the inoculation effect is present ($d = 0.463$), replicating the

findings from previous studies (Cook et al., 2017; Maertens et al., 2020; van der Linden et al., 2017), and also remains significant for at least one week, similar to what was shown in Study 1. Better than expected, we found that the inoculation effect can last up to at least one month without any booster intervention, with an effect compared to the misinformation-only control group of $d = 0.281$. These findings are in line with recent studies that suggest that inoculation effects may remain significant for at least two weeks (Banas & Rains, 2010) and that some inoculation effects may remain significant for up to 6 weeks (Ivanov et al., 2018), but not with other studies suggesting a full decay of the effect within 2 weeks (Zerback et al., 2021). Descriptive analyses also show that there is a decay in the inoculation effect over time, following a pattern similar to an exponential forgetting curve (Ebbinghaus, 1885; Murre & Dros, 2015).

Contrary to expectations, and despite motivation being a significant predictor of the outcome variable and the intervention successfully conferring long-term resistance to persuasion, the intervention did not have a significant effect on motivation. This shows that even without having an influence on motivational threat, it is possible that inoculation interventions can have long-lasting effects. This leads to important questions for the inoculation literature: it has long been assumed that a form of motivation or threat is an essential ingredient for inoculation interventions to be successful (Banas & Richards, 2017; Richards & Banas, 2018). As we do find that motivation was a predictor of the perceived scientific consensus after exposure to misinformation, it is clear that motivation does have a role. It may simply be that the warning message within the inoculation treatment was not strong enough to have an impact on motivation. Future research will need to further investigate the dynamics between motivation and inoculation intervention effectiveness, and inoculation scholars will need to consider whether an inoculation intervention needs to always successfully manipulate motivation or if there are cases where this may not be

necessary. Meanwhile, intervention designers may wish to consider testing multiple versions of their inoculation message to ensure that they have a maximal influence on motivation, and thereby potentially further strengthening the inoculation effect.

Looking at the underlying mechanisms, we find support for memory of the intervention as a major predictor of the inoculation effect, more so than other traditional variables studied in the inoculation theory literature such as *motivational* and *apprehensive threat* (Banas & Richards, 2017; Richards & Banas, 2018), *issue involvement* (i.e., how much do you engage with the topic), *post-inoculation talk* (i.e., do you talk about the inoculation content), and *attitude accessibility* (i.e., how often do you think about your own attitude towards the topic; Dillingham & Ivanov, 2016; Ivanov et al., 2012; Pfau et al., 2003, 2004, 2005). In addition, we tested an approximation of the memory-motivation model. We found that the longevity of the inoculation effect is mainly determined by objective memory rather than motivational threat, which is a new finding that was not found in the literature before, but that nevertheless motivation remains important for both memory strength and for resistance against persuasion, which is in line with the literature (Banas & Richards, 2017; Richards & Banas, 2018). The finding that memory of the intervention is an important factor is also compatible with the findings by Sanderson et al. (2021), who found that people's memory of the materials was a significant predictor of their sensitivity to a misinformation retraction in a continued influence paradigm.

The study also explored the potential of an inoculation "booster shot" by presenting people in the booster group the same inoculation message again after a week. Similar to a vaccine, this study provides evidence that cognitive immunity may be boosted with a second dose of the same inoculation intervention, mainly functioning by boosting memory of what was learned during the original intervention. Although we found no significant effect of the booster group compared to the single inoculation group—which we hypothesize to be

because the inoculation effect was still significant in the single inoculation group—we found a larger inoculation effect size at T3 ($d_{Booster}$ = 0.475 vs. $d_{NoBooster}$ = 0.281) as well as a significant effect of the booster on memory compared to a single inoculation (d = 0.741). This is in line with a recent study by Ivanov et al. (2018) that found that a repetition of the original inoculation message can serve as an effective booster.

We recommend that future research further explore the role of memory as a construct relevant to inoculation scholarship, and in particular pertaining to its long-term effectiveness. Moreover, as one limitation of this study is that we only looked at time points up to 1 month, we hypothesize that if longer time-intervals would have been used, we would have eventually seen a full decay of the inoculation effect in the single inoculation group, with a potential significant effect of the booster group in comparison to both the control and the single inoculation group. Another limitation of this study is that because all groups were separated for each time point to eliminate repeated testing effects, we were not able to test the memory-motivation model in its entirety as we could not calculate the paths between the different time points. Further scholarship could explore this further with a model that maps both the influence of motivation at T1 on memory at T1, and the inoculation effect at T3, in the same model. In the next two Chapters I will seek to replicate the findings of this study in two different inoculation paradigms: a broad, active, prophylactic, gamified intervention, and a broad, passive, prophylactic, video-based intervention.

Chapter 5

Gamified Inoculation

In this chapter, I present two studies (**Study 3–4**) where I explore the memory-motivation model in parallel to the previous chapter, but now in a gamified inoculation paradigm. In the first study (**Study 3**), published in the *Journal of Experimental Psychology: Applied*, I set up a new longitudinal experiment based on the *Bad News* paradigm. The purpose of this study is to show the feasibility of a longitudinal design based on an active, broad-spectrum, and prophylactic intervention, as well as show initial compatibility with the memory theory of inoculation. The three experiments featured in this study enable new insights into the memory theory of inoculation applied to gamified inoculation interventions by looking at 1) the short-term stability of inoculation (strong initial memory), 2) the long-term decay over time, and 3) the potential of repeated testing boosters to prevent decay. For the published version of the article, please visit the full publication at https://doi.org/10.1037/xap0000315. All datasets, analysis scripts, supplementary materials, and the preregistration are available on OSF repository at https://doi.org/10.17605/OSF.IO/2DTKB.

Afterwards, in **Study 4**, I zoom in into the mechanisms behind the decay, and create and test an additional booster treatment in the form of a shortened version of the original intervention. The results indicate that without an immediate posttest, the effect decays to a level of insignificance within 9 days after the intervention, demonstrating the importance of an immediate posttest for memory strengthening. The shortened *Bad News* booster intervention was successful at boosting memory, but was not sufficient to maintain the effect for 29 days. Using a range of traditional and new inoculation measures, I highlight the importance of memory for the inoculation effect longevity in gamified interventions, and find

new evidence for the validity of the memory-motivation model. The preregistration, survey materials, clean and raw datasets, and survey files are available on the OSF repository at https://osf.io/hwmge/?view_only=82bf2bc0f6ec4c5680e728cf5975244a.

Study 3

5.1 The Long-Term Effectiveness of Inoculation Against Misinformation: Gamified Interventions[27]

5.1.1 Abstract

This study investigates the long-term effectiveness of active psychological inoculation to build resistance against misinformation. Using three longitudinal experiments (two pre-registered), we tested the effectiveness of *Bad News*, a real-world intervention in which participants develop resistance against misinformation through exposure to misinformation techniques. In three experiments (N_{Exp1} = 151, N_{Exp2} = 194, N_{Exp3} = 170), participants played either *Bad News* (inoculation group) or *Tetris* (gamified control group) and rated the reliability of news headlines that either used a misinformation technique or not. We found that participants rate fake news as significantly less reliable after intervention. In Experiment 1, we assessed participants at regular intervals to explore the longevity of this effect and found that the inoculation effect remains stable for at least three months. With Experiment 2, we sought to replicate these findings without regular testing and found significant decay over a two-month time period so that the long-term inoculation effect was no longer significant. In Experiment 3, we replicated the inoculation effect and investigated whether long-term effects could be due to item-response memorisation or the fake-to-real ratio of items presented, but found that this is not the case. We discuss implications for inoculation theory and psychological research on misinformation.

[27] Study 3 has been published as *"Long-Term Effectiveness of Inoculation Against Misinformation: Three Longitudinal Experiments"* in the Journal of Experimental Psychology: Applied (Maertens et al., 2021). It was written in collaboration with Dr Jon Roozenbeek (University of Cambridge), Melisa Basol (University of Cambridge), and Professor Sander van der Linden (University of Cambridge). I am the sole first author on this work.

5.1.2 Introduction

Fake news can pose a serious threat to science, society, and democracy (Lewandowsky et al., 2017) with false content spreading faster and deeper on social networks than accurate or factual news (Petersen et al., 2019; Vosoughi et al., 2018). Although fake news may not usually constitute a majority of people's media diet (Allen et al., 2020), including during elections (Allcott & Gentzkow, 2017; Bovet & Makse, 2019; Grinberg et al., 2019), the risk can nonetheless be substantial. For example, the spread of false child abduction rumours on WhatsApp has led to deadly mob lynchings (Arun, 2019). Recent viral misinformation about the COVID-19 pandemic has led to the spread of dangerous health recommendations such as drinking bleach (Frenkel et al., 2020) and conspiracies about 5G networks worsening or causing COVID-19 symptoms have been associated with violent intentions (Jolley & Paterson, 2020) and contributed to people vandalising at least 50 phone masts in the UK alone (K. Chan et al., 2020). Accordingly, psychological research has seen a renewed interest in evaluating effective methods to counteract persuasion by (online) misinformation (Lazer et al., 2018).

Research on the social and cognitive correlates of belief in fake news has flourished, finding that although ideological motivations play a role in the perception and dissemination of misinformation (Grinberg et al., 2019; Guess et al., 2019; Jost et al., 2018; Swire, Berinsky, et al., 2017; van der Linden et al., 2020), higher cognitive ability and analytical thinking are generally associated with reduced belief in fake news (Bago et al., 2020; Bronstein et al., 2019; De keersmaecker & Roets, 2017; Pennycook & Rand, 2019, 2020; Swire, Berinsky, et al., 2017). Specifically, the finding that cognitive ability is strongly associated with susceptibility to misinformation opens up opportunities for the development of interventions. Accordingly, over the past years, researchers across disciplines have focused on creating solutions to effectively combat misinformation. One predominant approach

focuses on the efficacy of debunking and debiasing (Lewandowsky et al., 2012). Debunking, however, can be difficult, as fact-checks and corrections about contested issues may fail in light of (politically) motivated cognition (Flynn et al., 2017). Although the prevalence of the worldview backfire-effect is now increasingly debated (see Ecker, Lewandowsky, Fenton, et al., 2014; Ecker & Ang, 2019; Guess & Coppock, 2018; Wood & Porter, 2019), the continued influence effect (CIE) of misinformation can still limit the effectiveness of debunking techniques (Lewandowsky et al., 2012). Once exposed to a falsehood, it is difficult to correct, as people will often continue to rely on debunked information even when they acknowledge a correction (M.-P. S. Chan et al., 2017; Swire, Ecker, et al., 2017). Moreover, even when debunking is successful, regular exposure to misinformation can increase its perceived accuracy (Pennycook et al., 2018; Swire, Ecker, et al., 2017). Lastly, because fake news tends to spread faster and deeper than other types of news, fact-checkers continually remain behind the curve (Vosoughi et al., 2018).

As such, an attractive alternative approach to debunking is *prebunking*: protecting individuals against future persuasion attempts. Hornsey and Fielding (2017) propose a "jiu-jitsu" analogy of defence against persuasion attacks, which involves using the weight of an opponent against themselves. Inoculation follows a similar approach: by becoming familiar with persuasion techniques, people can protect themselves from being persuaded by misinformation.

Inoculation Theory

A Vaccine for Brainwash – William J. McGuire (1970, p. 36)

The "grandparent theory of resistance to attitude change" is inoculation theory (Eagly & Chaiken, 1993, p. 561). The process of inoculation follows a biomedical immunisation analogy, where exposure to a weakened strain of a pathogen triggers the production of antibodies to confer protection against future infection. In a similar fashion, inoculation

theory posits that people can build up cognitive resistance against unwanted persuasion attempts through "prebunking", i.e. by pre-emptively exposing people to weakened doses of persuasive arguments (Compton, 2013; McGuire, 1961, 1973; McGuire & Papageorgis, 1962). Over 60 years of research has shown that inoculation is among the most effective frameworks to help people resist persuasion attempts (Banas & Rains, 2010; Compton & Pfau, 2005). The inoculation procedure includes two components: *forewarning* and *refutational preemption*, which influence both cognitive and affective processes. Participants build up a set of skills to refute counterarguments and are made aware that their attitudes are vulnerable to more attacks in the future (creating a sense of *threat*; Compton & Pfau, 2005). The operationalisation of threat was traditionally left implicit ("inherent threat") and was theorised to be elicited by refutational preemption of counterarguments (Pfau et al., 1997), while the explicit forewarning ("extrinsic threat") was only introduced at a later stage (McGuire, 1964). More recent developments point towards an affective response to forewarning-induced threat which enhances resistance (Compton & Ivanov, 2014). Whether threat is a vital component for inoculation or not is actively debated (Banas & Rains, 2010; Banas & Richards, 2017; Compton, 2009). Originally it was argued that threat was essential (McGuire, 1964; McGuire & Papageorgis, 1962), especially to distinguish inoculation from two-sided messages (Compton & Pfau, 2005; Miller et al., 2013). Although some scholars have indeed demonstrated the importance of the role of threat (Compton & Ivanov, 2012; Richards & Banas, 2018), others have argued that this may not be a crucial component for conferring resistance to persuasion (Banas & Rains, 2010; Compton, 2009).

Although McGuire's own interpretation of the inoculation theory focused primarily on bolstering (existing) positive attitudes toward cultural truisms (e.g., brushing your teeth after a meal), contemporary inoculation scholarship now distinguishes between purely prophylactic and *therapeutic* inoculation approaches (Compton, 2019; van der Linden &

Roozenbeek, 2020). In fact, scholars have argued that the inoculation analogy should be "more instructive than prescriptive" (Compton, 2013, p. 233), and that "the therapeutic inoculation analogy can inspire a new generation of inoculation research" (Compton, 2019, p. 10). Just as therapeutic vaccines can still suppress infection by boosting the immune response, research has shown that people can also be inoculated against misinformation even when the message is not congenial to their prior attitudes, such as in the context of misinformation about climate change (Cook et al., 2017; Maertens et al., 2020; van der Linden et al., 2017). In a recent therapeutic intervention, Roozenbeek and van der Linden (2019) found that the largest inoculation effects were observed among those who were most susceptible to fake news prior to the intervention. Moreover, in an attempt to make inoculation theory scalable beyond specific issues, a second innovation has been a move away from argument-specific *narrow-spectrum* inoculations to *broad-spectrum* inoculations that focus on conferring resistance against a range of common *techniques* used in the production of misinformation (Basol et al., 2020; Cook et al., 2017; Roozenbeek, van der Linden, et al., 2020; Roozenbeek & van der Linden, 2019a, 2019b; van der Linden & Roozenbeek, 2020). In fact, McGuire (1961) himself hypothesised that one important factor in increasing the scope of protection was the notion of "active" rather than passive inoculation. In the active form of inoculation, participants have to generate their own "antibodies" or counterarguments. One example of active inoculation in the context of fake news is the *Bad News* game, a popular intervention which has been played by over a million people worldwide, and has been translated into more than 17 languages in collaboration with the UK Foreign and Commonwealth Office (Roozenbeek, van der Linden, et al., 2020).[28]

[28] The online game is free and publicly available at www.getbadnews.com.

Bad News Game

The *Bad News Game* is a real-world online intervention designed by Roozenbeek and van der Linden (2019) in collaboration with the Dutch media platform DROG, based on the principles of inoculation theory. In this free browser game, players enter a simulated social media environment and take on the role of a fake news producer. They design Twitter posts, news article headlines, and memes to gain popularity as a news publisher (see Figure 5.1.2.1 for an in-game screenshot). Players must gain followers while maintaining a sufficiently high level of credibility. If the credibility meter drops too low, the player loses, and the game ends. This way, the player is forced to think actively about how one can be deceived.

Often using a combination of humour and entertainment, the purpose of the intervention is to expose people to severely weakened doses of the techniques commonly used in the production of online misinformation. The game features six specific misinformation techniques known as DEPICT (the "six degrees of manipulation"), including ***D****iscrediting opponents* (e.g., creating a cloud of doubt around your opponent)*, appealing to* ***E****motion* (e.g., the use of outrage or highly emotive language to manipulate people), ***P****olarizing audiences* (e.g., using hot-button issues to drive a wedge between two groups), ***I****mpersonation* (e.g., misusing the identity of politicians, experts, or celebrities online), *floating* ***C****onspiracy theories* (e.g., casting doubt on mainstream narratives by providing an attractive story in which a small sinister group of people is responsible for doing harm to many), and ***T****rolling* (e.g., eliciting reactions from people by provoking them online). See Roozenbeek and van der Linden (2019) and van der Linden and Roozenbeek (2020) for a detailed background and overview of these techniques.

The game was designed to incorporate the components necessary for inoculation (Roozenbeek, van der Linden, et al., 2020). During gameplay, the player is required to imagine how misinformation techniques could be refuted, which serves as the active

refutational element of the inoculation. The scenarios were designed to provide participants with a slightly uncomfortable feeling (as they are responsible for creating and sharing fake news), thus eliciting a sense of threat.[29] As opposed to being issue-based, threat in broad-spectrum inoculation is understood as making the dangers of the spread of fake news salient (by exposure to weakened doses). In fact, the game scenarios themselves incorporate a strong *forewarning* component to foreshadow how fake news can have damaging consequences. For example, participants are explicitly warned about how emotions can be exploited in the media or that "conspiracy theories can be a great way of spreading disinformation". Threat is also elicited directly by attacks from other simulated "users" through a wide range of social media content. For example, when players choose options that are not in line with the purpose of the game, motivation to do so is boosted by issuing a warning that elevates the threat level; "Whoops, we're running into a bit of a problem, some 'fact-checker' has taken notice … seriously you need to have a look at this."

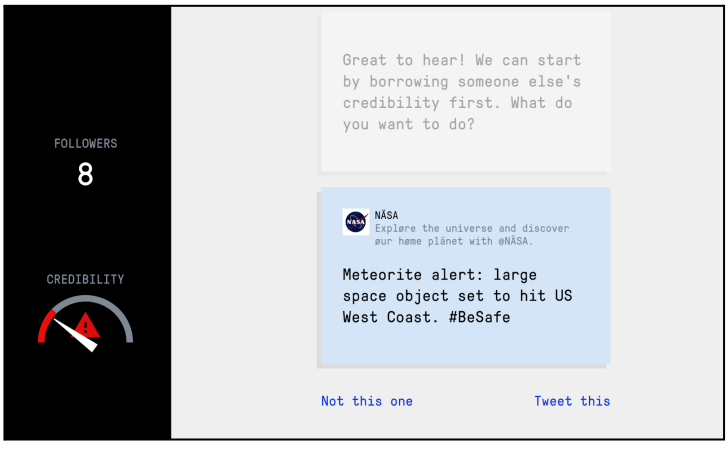

Figure 5.1.2.1. Screenshot of *Bad News* Game Environment.

In their original study, Roozenbeek and van der Linden (2019) used a within-subjects design to evaluate the efficacy of the *Bad News Game* as a "broad-spectrum vaccine" against

[29] A prior study analysed open-ended responses as part of a post-gameplay survey and found that the game elicits more (negative) affect compared to a control group (Roozenbeek & van der Linden, 2019a).

fake news. In their study, about $N = 15{,}000$ participants rated the reliability of several fake and real news items (in the form of fictitious Twitter posts) pre and post gameplay. Notably, these were different items than people were trained on in the game (i.e., a refutational-*different* approach to inoculation). The researchers found that while the *fake news* items corresponding to the misinformation techniques were rated as significantly less reliable after playing the game ($d_{average} = -0.52$), people did not meaningfully adjust their ratings of the *real news* items. Subsequent experiments have shown that the *Bad News* intervention also boosts *confidence* in people's truth-discernment abilities (Basol et al., 2020) and that the inoculation effect generalises across different cultural contexts (Roozenbeek, van der Linden, et al., 2020). Yet, importantly, nothing is currently known about the duration of the inoculation effect conferred through *Bad News*, which is a crucial factor in not only determining the long-term efficacy of the intervention, but also in advancing our understanding of inoculation theory and immunity to persuasion.

Longevity

Although the effectiveness of inoculation theory has been well established, research on its long-term effectiveness remains an area with many open questions (Banas & Rains, 2010). Importantly, the rate of the decay of the treatment effect of the *Bad News* intervention is therefore not only of practical utility: it is also a question of high theoretical significance. However, there is currently no coherent theoretical framework that accurately predicts a specific decay[30] function of resistance to persuasion.

Although McGuire (1964) initially argued that a delay of a few days between inoculation and attack is needed in order to build up sufficient "mental antibodies", the more consistent finding points towards the opposite: decay in the inoculation effect over time

[30] Throughout this paper, we use the word *decay* as a theory-neutral description of the decrease in the inoculation effect over time, unless otherwise specified. The term 'decay' is often used in inoculation research in general terms without making claims about whether the decay is *due to* the mere passage of time, or related to memory function. In contrast, in memory research, *decay* often refers to *trace decay*, a specific memory theory (Brown, 1958).

(Banas & Rains, 2010; Ivanov, 2012). Research indicates that this decay might be slower than the decay found when using other methods such as narrative messaging (Niederdeppe et al., 2015) or consensus messaging (Maertens et al., 2020). A recent study into the effectiveness of a digital media literacy intervention found over 50% decay of the effect over three weeks (Guess et al., 2020). While some studies show that the inoculation effect decays within two weeks (Zerback et al., 2021), other findings suggest that inoculation effects can remain detectable for up to six weeks (Pfau et al., 2004, 2006). A meta-analysis suggested an unchanged (stable) effect with a duration of at least two weeks, followed by a decay of the effect after this period of stability. The most recent study to systematically explore inoculation decay at multiple time points found that decay started between four to six weeks after intervention (Ivanov et al., 2018). The longest retention figures suggest that some inoculation effects could sustain over a period of 33 weeks (Pfau et al, 1992). Yet, it remains unclear whether the inoculation decay function is continuous or intermittent; linear, curvilinear or exponential; and if the decay function can take different forms under specific circumstances (Banas & Rains, 2010; Compton & Pfau, 2005; Ivanov, 2017).

The *Bad News* intervention is a particularly interesting test case as active inoculation is meant to stimulate analytical thinking and strengthen linkages between nodes in associative memory networks, which are thought to both facilitate resistance to persuasion and improve the longevity of the inoculation effect (Banas & Rains, 2010; Pfau, Ivanov, et al., 2005; Pfau, Tusing, et al., 1997). For example, based on network models of memory (Anderson, 1983; Forgas, 2001), Pfau et al. (2005) theorised that resistance to persuasion might nest itself in long-term memory networks. They argued that an attitude could be represented as an associative memory network with cognitive and affective nodes. Based on Petty et al. (1994), they posited that a more dense network would be more resistant to change. Using concept mapping as a method to represent mental structures, they found increases in relevant nodes

and linkages after an inoculation message, which in turn led to more resistance to persuasion attacks at a later date (Pfau et al., 2005).[31]

More generally, to counter the decay of the inoculation effect, evidence has been found for the effectiveness of *booster treatments* (Ivanov et al., 2018). It is theorised that similar to biomedical inoculations, a regular "booster shot" may be needed to top up the cognitive immune system (McGuire, 1961). Examples of booster messages include a weakened attack message, a repetition of the inoculation procedure (in full or shortened form), or a new warning message to elicit a fresh sense of threat (Ivanov et al., 2018). While evidence on the effectiveness of booster treatments is mixed (Compton & Pfau, 2005; Ivanov et al., 2009; Pfau, 1992), the general conclusion is that boosters work when administered in the right form at the right time (Ivanov, 2012; Ivanov et al., 2018; McGuire, 1961; Pfau et al., 2004). Further, the concept of booster treatments could be interpreted through a memory lens as *relearning*, leading to stronger memory representations (Ebbinghaus, 1885; Murre & Dros, 2015). Ivanov et al. (2018, p. 661) stress that "the book on boosters is not ready to be closed" and that we need to "reignite the research interest in inoculation booster messages."

The Present Research

Since prior evaluations of the *Bad News* intervention were not pre-registered, the first goal of the current study was to replicate the original effect of the *Bad News* intervention in a pre-registered experimental study with a larger battery of fake news test items. Based on prior work, we expected to replicate the main effect of the intervention.

> **H_1**: *On average, participants in the inoculation group rate fake news (post - pre) as significantly less reliable compared to (post - pre) ratings of the same items in the control group.*

[31] With concept mapping participants have to draw a map similar to a mind map. Participants are asked to think about and write down everything related to a central topic (i.e., the inoculation topic). The different nodes (circles with concepts) and the links between the nodes they draw, count as the density of the memory network.

Given a paucity of research on the longevity of inoculation interventions, we advance the literature by measuring the effectiveness of the inoculation intervention over time. Although there is no clear theory that would predict the longevity of the inoculation effect, based on the work reviewed above, we can conclude that the inoculation effect decays over time. Based on the meta-analysis finding of decay starting at some point after two weeks (Banas & Rains, 2010), and a recent study showing decay setting in between four to six weeks (Ivanov et al., 2018), we hypothesise that the decay process should happen within the timeframe of two months. This led to our second (decay) hypothesis.

H_2: *After two months, participants show a significant decrease in the inoculation effect.*

5.1.3 Methods

Design and Procedure

For Experiment 1 and Experiment 2, we utilised a randomised pretest-posttest design (Campbell, 1957; Huck & McLean, 1975). Participants were randomly allocated to either the inoculation group or the control group. In Experiment 3, all participants received the inoculation intervention. In all three experiments, participants started with a pretest survey. In this survey, participants had to judge the reliability of news items (21 in Exp 1-2, seven in Exp 3) that were either factual news headlines (three in Exp 1-2, one in Exp 3) or headlines featuring a misinformation technique (18 in Exp 1-2, six in Exp 3). All participants had to rate the reliability of the news headlines on a Likert scale from 1 (very unreliable) to 7 (very reliable). All items were presented in random order. After rating the news items at pretest (T1), participants were asked to complete *Bad News* (inoculation group), or to play ~15 minutes of *Tetris* (control group). The 15-minute time slot was chosen to match the completion time of the *Bad News* game. After the intervention, participants were asked to rate

the reliability of the same headlines again (T2). In Experiment 1 and Experiment 3, participants were directed to a demographics survey after the posttest, and answered questions about their year of birth, gender, political affiliation (from 1-7, very left-wing to very right-wing), country of residence, first language, social media usage (from 1-5, never to daily), and had to respond to a single-item cognitive reflection test: *"A ball and a bat cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?"* (Frederick, 2005). All participants received a debriefing at the end.

The same participants were then contacted again (T3) to participate in the follow-up where they had to rate the reliability of the same (Exp 1-2) or different (Exp 3) news headlines. T3 took place one week after the initial test date for Experiments 1 and 3, and two months after the initial test date in Experiment 2. Participants then received an extended debriefing. Participants in Experiment 1 were asked to participate in another follow-up four weeks after the second posttest (T4), and then again eight weeks after this (three months after the initial test date; T5). For consistency, all items were kept the same throughout the different test dates in Experiment 1 and Experiment 2, while Experiment 3 specifically investigated whether using the same (versus different) items produces a confound. See Figure 5.1.3.1 for a visualisation of the study flowchart depicting the experiments' design.

All three experiments were approved by the Cambridge Psychology Research Ethics Committee (ref. PRE.2018.085, PRE.2019.103). Our pre-registered hypotheses can be found on the AsPredicted platform (Experiment 1: https://aspredicted.org/xn6qy.pdf, Experiment 3: https://aspredicted.org/ka2at.pdf). All deviations from the original pre-registration can be found in Supplementary Declaration S1. All datasets, measurement items, and analysis scripts for Experiment 1, 2, and 3 are publicly available on our OSF repository: https://doi.org/10.17605/OSF.IO/2DTKB.

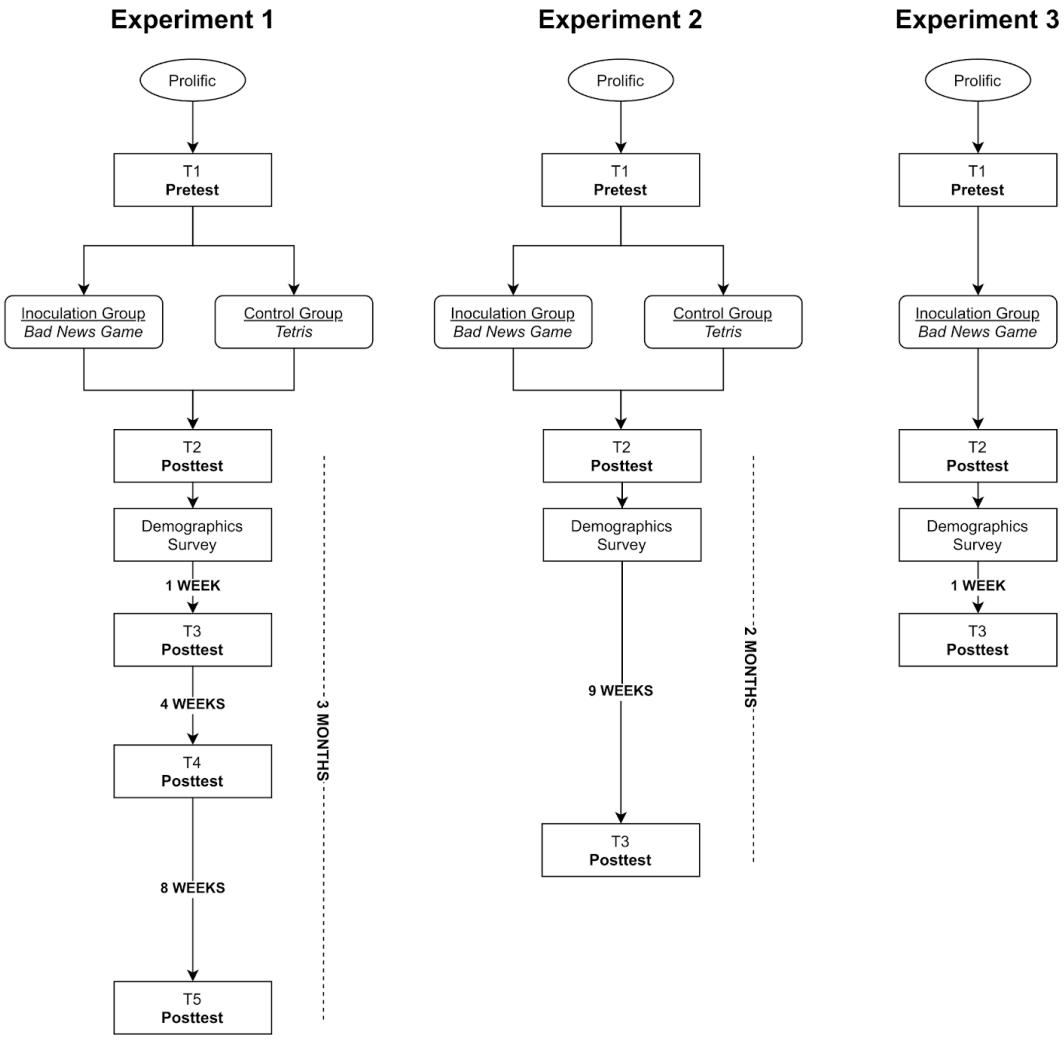

Figure 5.1.3.1. Overview of Experiment flowchart of Experiment 1, 2, and 3.

Measures

In the initial *Bad News* study by Roozenbeek and van der Linden (2019), only one fake news item per misinformation category was evaluated by the participants, and the items were not randomised. To make the measure stronger and to avoid the possibility of item-effects (Roozenbeek, Maertens, et al., 2021), we developed three manipulative news headlines per misinformation technique, plus three factual 'real news'[32] control headlines. Importantly, whilst modelled after real-world events, the test items were a) fictional and b) different from those used in the inoculation training itself. Participants could, therefore, not

[32] These were not of primary interest here but included to remain consistent with the approach of Roozenbeek and van der Linden (2019). An overview of all items can be found in Supplementary Information S1.

just rely on recognition but were required to detect the misinformation strategy in a new setting. Using fictional items also maximises experimental control over isolating the manipulation techniques and avoids familiarity confounds with real fake news (Roozenbeek, Maertens, et al., 2020; Roozenbeek, van der Linden, et al., 2020). The items were designed based on the most commonly used misinformation techniques (using the definitions provided in the original study; Roozenbeek & van der Linden, 2019a). An example of a real news item would be "*Physical fitness keeps your brain in good shape*," posted by "*@PsychologyToday*." This item does not contain any misleading strategy, while the fake news items each have their own misinformation element. Examples include "*Scientists discovered solution to greenhouse effect years ago but aren't allowed to publish it, report claims*" (conspiracy) and "*New study shows right-wing people lie much more than left-wing people*" (polarization). See Figure 5.1.3.2 for an example of how the items were presented. Consistent with Basol et al. (2020) all analyses were done with an aggregate index of all fake news items (M = 3.39, SD = 0.81, Cronbach's α = 0.83).[33]

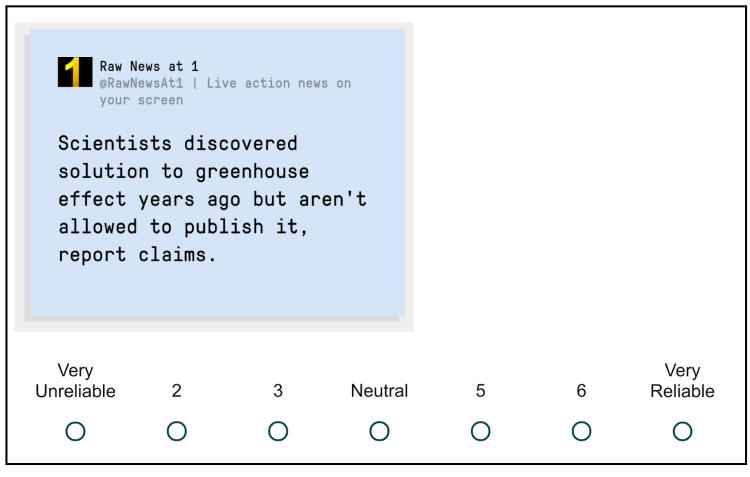

Figure 5.1.3.2. Example test item using the conspiracy technique.

[33] This is a deviation from the pre-registration to prevent multiple testing artefacts and increase internal consistency of the measurement. For these reasons, we only use the combined fake news index (18 items) and neither analyse subcategories (3 items per category only) nor real news (3 items only). A scree plot provides evidence for the unidimensionality of the 18 fake news items (only one factor with eigenvalue > 1; see Supplementary Figure S5). Nonetheless, results per subcategory are provided in Supplementary Tables S3-S8 and real news items are plotted in Supplementary Figures S1-S4.

Empirical Strategy

We follow our pre-registered empirical strategy with one change. To help make a clearer distinction between our two hypotheses and make the coefficients interpretable, we separated the analyses for both Experiment 1 and Experiment 2 into a standard analysis of covariance (ANCOVA) for Hypothesis 1 and a repeated-measures analysis of covariance (rmANCOVA) for Hypothesis 2.[34]

To test $\mathbf{H_1}$, we performed an ANCOVA with *intervention* as a between-subjects factor with two levels: *Bad News* (Inoculation Group) vs. *Tetris* (Control Group), and posttest reliability rating (T2/T3/T4/T5) as the dependent variable. As our primary concern was measuring the difference between the control group and the inoculation group after intervention without confounds by potential group differences, we modelled the pretest (T1) measure as a covariate (Coppock, 2019).

To test $\mathbf{H_2}$, we used the same ANCOVA but with the repeated measures variable time added as a within-subjects factor with two levels: T2 (day 1, after intervention) and T3, T4, or T5, and the interaction between time and the intervention.

In Experiments 1 and 2, the fake-to-real item ratio (18:3) was unbalanced, and the same item sets were used for each assessment. We, therefore, added Experiment 3, a replication of Experiment 1 (up to T3) with different item sets and a balanced item ratio (6:6). To test the validity of the hypothesis tests after taking into account these potential confounds, we followed the pre-registered empirical strategy of using within and between-subject *t*-tests to compare the inoculation effects within and between the two experiments.

[34] The pre-registration only mentions the rmANCOVA model. As this model is confounded with the time variable, and the pure inoculation effect hypothesis H_1 is unrelated to the time variable, we decided to split these analyses. All deviations from the pre-registration can be found in Supplementary Declaration S1.

5.1.4 Experiment 1

Method

Participants and Sample Composition

In the original large-sample *Bad News* study, an average inoculation effect size of d = 0.52 was found for the fake news scale (Roozenbeek & van der Linden, 2019a). A power analysis with an effect size of d = 0.52, a significance level of .05, .80 power, and taking in account potential participant attrition of 20% over the test period, indicated a required sample size of 148 participants. Based on this, we recruited a total of 151 participants.

Participants were recruited through the online platform Prolific (https://prolific.ac/) and were rewarded 2.35 GBP if they participated in both the initial experiment (T1/T2) and the follow-up one week later (T3). They were unaware of the later follow-ups (T4, T5), but were later invited to participate for 0.25 GBP per bonus follow-up.[35] All participants gave informed consent before participation. The sample consisted of 151 participants (81 control, 70 intervention), 52% identifying as female, slightly skewed towards younger age (19-66, *Mdn* = 28). The sample was culturally diverse with 23 different countries of which the largest are the United Kingdom (29%), Italy (17%), and Poland (15%). Political ideology was skewed towards left-wing (49% left-wing, 19% right-wing; M = 3.50, SD = 1.25), and 50% with a higher education level diploma. For a complete overview of the sample demographics separated by the T1 sample and the complete-cases T5 sample, see Supplementary Table S1.

As pre-registered and accounted for in our power analysis, we only use the dataset with complete cases (N_{all} = 151, $N_{complete}$ = 118, attrition = 22%) for our hypothesis tests. We did this to have comparable results between the different test dates. An attrition analysis indicated that no specific demographic factors (e.g., age, education, ideology) could predict

[35] The bonus follow-ups were not pre-registered because we originally did not plan them. See Supplementary Declaration S1 for an overview of all deviations from the pre-registered procedure.

the pattern of attrition (see Supplementary Table S9), and pretest reliability could not predict completeness (b = -0.04, OR = 0.96, 95% CI_{OR} [0.59, 1.54], p = 0.87).

Results

Inoculation Effect

The average inoculation effect was significantly stronger in the treatment vs. control group, manifested by a significant main effect of *intervention* in our ANCOVA analysis (F(1, 114) = 29.86, p < .001, η^2_p = 0.21, η^2 = 0.07). Difference-in-differences analysis ($M_{diffT2T1,control}$ = -0.08, $SD_{diffT2T1,control}$ = 0.34; $M_{diffT2T1,inoculation}$ = -0.61, $SD_{diffT2T1,inoculation}$ = 0.66) using a post-hoc t-test indicated a significant mean difference of $M_{diff\text{-}in\text{-}diffs}$ = -0.52[36] (t(84) = -5.41, p < .001, 95% CI [-0.72, -0.33], d = -1.00), indicating that participants who played the *Bad News Game* rated the fake news items as significantly less reliable than those who played *Tetris*.[37]

Decay: One Week

To test H_2, we first used the same ANCOVA model to verify if the inoculation effect was still significant, but now with the one-week-later posttest (T3) as the dependent variable. We found that the inoculation effect was still significant (F(1, 114) = 23.91, p < .001, η^2_p = 0.17, η^2 = 0.07; see Figure 5.1.4.1, panel A, T3). Using an rmANCOVA with both the immediate posttest (T2) and the posttest one week later (T3) to test the decay hypothesis, no indication for an interaction effect between time and intervention was found (F(1, 229) = 0.02, p = .88, η^2_p = 0.00, η^2 = 0.00).

Decay: Five Weeks

To extend our original hypothesis, we looked at potential decay five weeks after the intervention (T4). Using the same ANCOVA, we found that the inoculation effect was still significant (F(1, 114) = 27.68, p < .001, η^2_p = 0.20, η^2 = 0.10; see Figure 5.1.4.1, panel A,

[36] The difference (0.61 - 0.08) was 0.52 and not 0.53 due to rounding. Raw values (no rounding) were used for all calculations.

[37] All raw means and confidence intervals for Experiment 1 are available in Supplementary Table S3-S6.

T4). When looking at the rmANCOVA, we again found no significant interaction effect between time and intervention ($F(1, 229) = 0.44$, $p = .51$, $\eta^2_p = 0.00$, $\eta^2 = 0.00$).

Decay: 13 Weeks

We once more extend our scope with a follow-up three months after the initial intervention (T5). The inoculation effect still remained significant ($F(1, 114) = 13.27$, $p < .001$, $\eta^2_p = 0.10$, $\eta^2 = 0.06$; see Figure 5.1.4.1, panel A, T5). The rmANCOVA result showed no significant interaction effect between intervention and time ($F(1, 229) < 0.01$, $p = .98$, $\eta^2_p = 0.00$, $\eta^2 = 0.00$). Using a post-hoc *t*-test difference-in-differences calculation, we found an inoculation effect retention of 100%[38] between T2 and T5 ($M_{\text{diffT5T2,control}} = -0.19$, $SD_{\text{diffT5T2,control}} = 0.81$; $M_{\text{diffT5T2,inoculation}} = -0.19$, $SD = 0.69$; $M_{\text{diffT5T2,inoculation}} = 0.00$, $t(114) = 0.00$, $p > .999$, 95% CI [-0.27, 0.27], $d = 0.00$). See Figure 5.1.4.1, panel B, for a summarising density plot for each test date.

[38] Retention formula: 1-((Control T5 - Control T1) - (Inoculation T5 - Inoculation T1) / ((Control T2 - Control T1) - (Inoculation T2 - Inoculation T1)).

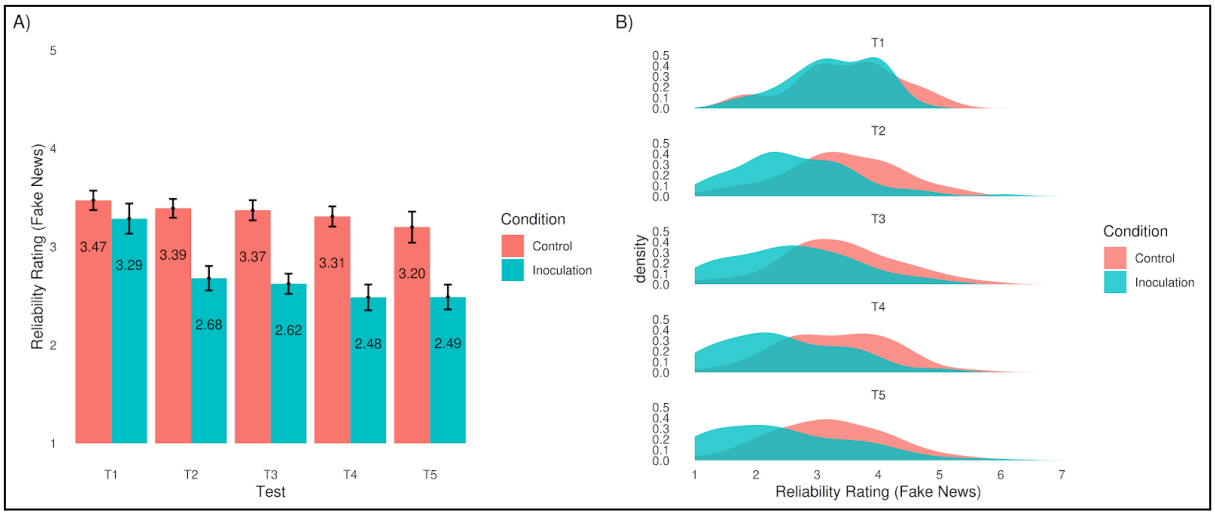

Figure 5.1.4.1.[39] Reliability ratings of fake news items, separated by time and condition in Experiment 1. Panel A: average reliability rating of fake news items over time. Panel B: density plots of these results. T1 = pretest. T2 = posttest (0 weeks). T3 = posttest (1 week). T4 = posttest (5 weeks). T5 = posttest (13 weeks). N = 118. Error bars represent 95% confidence intervals.

Exploratory

For exploratory purposes, we analysed and visualised a range of extra variables and examined the robustness of the inoculation effect when controlling for individual differences.[40] Using a linear regression model with the T2-T1 difference score as outcome variable we found that the inoculation effect remained significant, $t(106) = -5.29$, $p < .001$, $\beta = -0.46$, 95% CI [-0.66, -0.26], after controlling for gender, age, country of residence, education level, political ideology, social media use, Twitter use, cognitive reflection test performance, and primary language, consistent with previous results by Roozenbeek and van der Linden (2019).

[39] Alternative plots for Experiment 1 are available in Supplementary Figure S1-S2.
[40] See Supplementary Table S10 for linear model estimates and Supplementary Analysis S1 for visual plots.

Discussion

The lack of decay observed in Experiment 1 is contrary to our expectations. We hypothesise that the repeated tests might have confounded the result as they could function as booster sessions or simply testing effects. The regular exposure to weakened doses of each fake news technique (in this case, the measurement items) could serve as reminders that reinstate the inoculation effect. Ivanov et al. (2018) found that for longer time intervals booster messages can indeed prolong the inoculation effect, leading to significant inoculation effects for at least six weeks. These findings can be related to memory research, where a robust literature shows the memory-strengthening impact of repeated testing (Karpicke & Roediger, 2008; Linton, 1975; Roediger & Karpicke, 2006a, 2006b). By repeatedly requiring people to recall what they learned, they relearn these lessons (Nader & Hardt, 2009), potentially leading to an increase in inoculation effect stability over time.

To investigate the inoculation effect retention while eliminating the possibility of learning or boosting effects, we re-run the experiment, removing all intermediate tests to allow for a full two-month decay period without follow-ups (Experiment 2).

5.1.5 Experiment 2

Method

Design and Participants

In a parallel research project, a similar experiment was conducted with the same test items, but without any follow-up after the initial test date. We decided to leverage this opportunity to re-examine the potential for decay. We re-contacted this study's participants two months after their initial test and asked them to participate in an unexpected bonus follow-up, which functions as our Experiment 2. Importantly, the main procedures and

reliability measures are the same between Experiment 1 and Experiment 2, with one crucial difference: the different time interval between T2 and T3 (two months instead of one week).[41]

In total, $N = 194$ participants (107 control, 87 inoculation) were recruited through Prolific. In the unexpected follow-up two months later (T3), the number of participants was reduced to 110 (56 control, 54 inoculation), a 57% retention of the initial sample. In terms of the sample composition,[42] 57% were male, skewed towards younger age (18-44, modal bracket = 18-24). Political ideology was skewed towards left-wing (59% left-wing, 21% right-wing; $M = 2.40$, $SD = 1.34$), and 26% had a higher education level diploma (modal bracket = *High school diploma*, 51%).

We used the same empirical strategy as in Experiment 1. Similarly, as in Experiment 1, all hypothesis tests were done using complete cases ($N_{all} = 194$, $N_{complete} = 110$, attrition = 43%) and on the 18-item fake news scale ($M = 3.20$, $SD = 0.85$, $\alpha = 0.83$).[43]

Results

Inoculation Effect

An inoculation effect was found on the initial test date (T2) for the inoculation group ($F(1, 106) = 11.65$, $p < .001$, $\eta^2_p = 0.10$, $\eta^2 = 0.03$). A post-hoc *t*-test for difference-in-differences analysis ($M_{diffT2T1,control} = -0.08$, $SD_{diffT2T1,control} = 0.32$; $M_{diffT2T1,inoculation} = -0.50$, $SD_{diffT2T1,inoculation} = 0.80$) showed that this difference was a significant decrease in reliability ratings for the inoculation group compared to the control group ($M_{diff\text{-}in\text{-}diffs} = -0.42$, $t(69) = -3.62$, $p < .001$, 95% CI [-0.19, -0.65], $d = -0.69$).[44] See Figure 5.1.5.1, panel A, T1 and T2 for a visual comparison.

[41] One other notable difference is that two out of three real news items were actually different in Experiment 2 (see Supplementary Information S1 for a comparison).
[42] Note that, as participation in the demographics survey was optional in this experiment and conducted at a different date, education level, age, and gender were only answered by 39% of participants.
[43] As 61% of participants did not complete the extra demographics survey, we performed neither attrition analyses nor exploratory analyses for Experiment 2. However, model estimates and visual plotting of these analyses in Experiment 1 can be found in Supplementary Table S10 and Supplementary Analysis S1, respectively.
[44] All raw means and confidence intervals for Experiment 2 are available in Supplementary Table S7-S8.

Decay: Nine Weeks

Using the same analyses as for Experiment 1, we found no inoculation effect for the standalone ANCOVA ($F(1, 106) = 2.17$, $p = .14$, $\eta^2_p = 0.02$, $\eta^2 = 0.01$), and no interaction effect between time and intervention for the rmANCOVA ($F(1, 213) = 2.18$, $p = .14$, $\eta^2_p = 0.01$, $\eta^2 = 0.00$). Using a post-hoc *t*-test difference-in-differences analysis we found a non-significant inoculation effect retention of 36%[45] between T2 and T3 ($M_{diffT3T2,control} = -0.01$, $SD_{diffT3T2,control} = 0.78$; $M_{diffT3T2,inoculation} = 0.26$, $SD_{diffT3T2,inoculation} = 0.75$; $M_{diff\text{-}in\text{-}diffs} = 0.27$, $t(108) = 1.85$, $p = .07$, 95% CI [-0.56, 0.02], $d = -0.35$).

Plotting these results pointed towards a partial decay hypothesis (see Figure 5.1.5.1, Panel A, T1-T3). We found further evidence for this by visually analysing the distribution of the reliability ratings in the inoculation group (see Figure 5.1.5.1, Panel B, inoculation group), manifested by a dent in the plot indicating that some reverted to baseline while for others the inoculation benefits persisted.

[45] Retention formula: 1-((Control T3 - Control T1) - (Inoculation T3 - Inoculation T1) / ((Control T2 - Control T1) - (Inoculation T2 - Inoculation T1)).

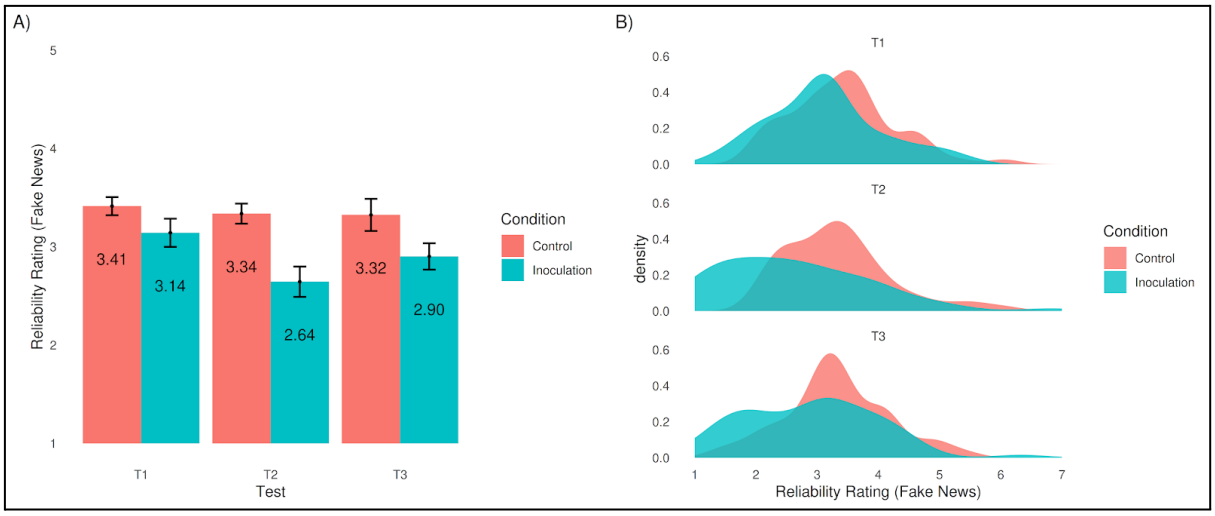

Figure 5.1.5.1.[46] Reliability ratings of fake news items, separated by time and condition in Experiment 2. Panel A: plot of average fake news reliability ratings. Panel B: density plots of the same results. T1 = pretest. T2 = posttest (0 months). T3 = posttest (2 months). $N = 110$. Error bars represent 95% confidence intervals.

Discussion

In Experiment 2, we eliminated the confound of repeated measurement by removing all follow-ups between the direct posttest and the posttest two months later. In line with the original hypothesis, we find that the inoculation effect indeed decays over the course of two months, rendering the effect no longer significant. The analyses also show that the decay is only partial, with density plots suggesting that the effect might still linger on for some participants. A final set of concerns left unanswered by the previous two experiments is whether confounds are introduced because of (a) the unbalanced fake-to-real ratio (18:3) of the presented news items (Aird et al., 2018) and (b) the fact that the same items were used at each follow-up, which may lead to item-response memorisation effects. Experiment 3 aims to rule out these alternative explanations.

[46] Alternative plots for Experiment 2 are available in Supplementary Figures S3-S4.

5.1.6 Experiment 3

Method

Design and Participants

In Experiment 3, we explored whether the sustained effects in Experiment 1 could be due to either the memorisation of responses to the items (all items were the same for each test date) or due to the skewed ratio (18:3) of fake-to-real news items (Aird et al., 2018). To accomplish this, we designed an experiment that was identical to Experiment 1 (up to T3, the first follow-up) but changed both the item set and fake-to-real ratio for the follow-up measure. In this pre-registered experiment[47], we omitted the control group, as we wanted to maximise power and because our core comparison of interest was the inoculation group. This design allowed us to compare the results of Experiment 3 to those of Experiment 1, to find out whether the T3 results are the same now that two confounds (item set repetition and fake-to-real ratio) have been eliminated. See Figure 5.1.6.1 for a comparison of the two experimental designs, and Supplementary Repository S1 for the precise item sets.

We conducted a power analysis with power = 0.80, $\alpha = 0.05$, $d = 0.45$ (SESOI), expected attrition = 10%, and $N_{Exp1} = 70$. We recruited 100 participants from Prolific. Participants in any previous *Bad News* experiments were barred from participation. We followed the same data cleaning procedures as for Experiment 1. Thirteen people dropped out for T3, making the final sample $N_{Exp3} = 87$. Our final sample was younger (*Mdn* = 22, 84% between 18-29), predominantly male (75% male, 23% female), more left-wing ($M = 3.45$, $SD = 1.41$), educated (45% with higher education diploma), and most participants came from Poland (29%) or Portugal (28%).

[47] https://aspredicted.org/ka2at.pdf. Any deviations can be found in Supplementary Declaration S1.

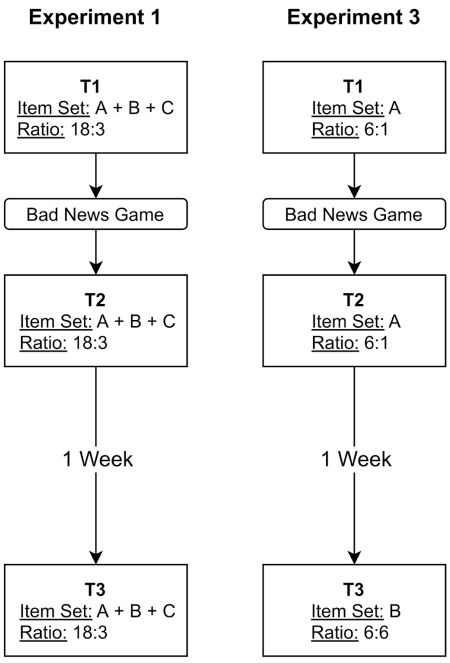

Figure 5.1.6.1. Flowcharts of Experiment 1 (up to T3, excluding control group) and Experiment 3, with item set information. Item sets are news sets with six fake news items and one to six real news items each. Ratio refers to the *fake-to-real ratio* of the items presented.

Results

Within-Group

As pre-registered, we first looked at whether the inoculation effect is present for each time point. When comparing T2 (M = 2.83, SD = 1.09) to T1 (M = 3.48, SD = 0.88), we found a significant negative effect with $M_{\text{diff,T2T1}}$ = -0.65, 95% CI_M [-0.84, -0.46], $t(86)$ = -6.70, $p < .001$, d = -0.72, 95% CI_d [-0.95, -0.48]. This effect shows that a medium-to-large baseline effect was established using the same item set (Set A).

We also compared T3 (M = 2.79, SD = 0.98) to T1 (M = 3.48, SD = 0.88), and found a near-identical significant effect with $M_{\text{diff,T3T1}}$ = -0.70, 95% CI_M [-0.90, -0.50], $t(86)$ = -6.87 = $p < .001$, d = -0.74, 95% CI_d [-0.97, -0.50]. We thus found a significant medium-to-large effect of the inoculation intervention using Set B, indicating that the intervention was effective despite using a different item set and after equalising the fake-to-real ratio. See

Supplementary Tables S12 and S13 for an overview of the raw means and difference scores for each time point and each item set.

Between-Groups

As pre-registered, the next step in our decision tree was to compare the within-group difference scores between both experiments, to explore if, despite the inoculation effect remaining significant, the altered experiment design influenced the treatment effect. We first looked at the T2-T1 difference in Exp 3 ($M_{\text{diff,T2T1}}$ = -0.65, SE = 0.10) compared to Exp 1 ($M_{\text{diff,T2T1}}$ = -0.68, SE = 0.10), and found no significant difference between the two groups with $M_{\text{diff-in-diffs}}$ = 0.03, 95% CI_M [-0.24, 0.31], $t(153) = 0.23$, $p = .82$, d = -0.04, 95% CI_d [-0.28, 0.35]. This difference was also statistically equivalent to zero ($t(153) = -2.59$, p = .005);[48] we could therefore conclude that the baseline effect was the same between the two samples.

We then compared T3-T1 difference in Exp 3 ($M_{\text{diff,T3T1}}$ = -0.70, SE = 0.10) to Exp 1 ($M_{\text{diff,T3T1}}$ = -0.91, SE = 0.14), and found no significant difference between the two experiments with $M_{\text{diff-in-diffs}}$ = 0.21, 95% CI_M [-0.12, 0.55], $t(134) = 1.25$, $p = .21$, d = -0.20, 95% CI_d [-0.11, 0.52]. Although this effect was not significant, it was not statistically equivalent to zero at the traditional α level ($t(133) = 1.53$, $p = 0.06$). These findings indicated that there was no significant increase in reliability ratings of fake news by changing the experimental design for the T3 follow-up one week later (although a small increase could not be ruled out).

Finally, looking at the T3-T2 difference in Exp 3 ($M_{\text{diff,T3T2}}$ = -0.06, SE = 0.10) compared to Exp 1 ($M_{\text{diff,T3T2}}$ = -0.23, SE = 0.11), we did not find a significant difference with $M_{\text{diff-in-diffs}}$ = 0.18, 95% CI_M [-0.11, -0.47], $t(152) = 1.21$, $p = 0.23$, $d = 0.19$, CI_d [-0.12, 0.51]. Equally, although the comparison was not significantly different, it was not statistically

[48] We used Two One-Sided Tests (TOST) Equivalence Testing using the TOSTER package in R with α = 0.05 and as Smallest Effect Size of Interest (SESOI) d = (-)0.45.

equivalent to zero ($t(152) = -1.60$, $p = 0.06$). This indicated that the inoculation retention over a one-week period was similar between the two experimental setups, thereby finding no evidence for item ratio or item set specific retention effects. See Figure 5.1.6.2 (Panel A) for a bar chart comparing the two experiments.

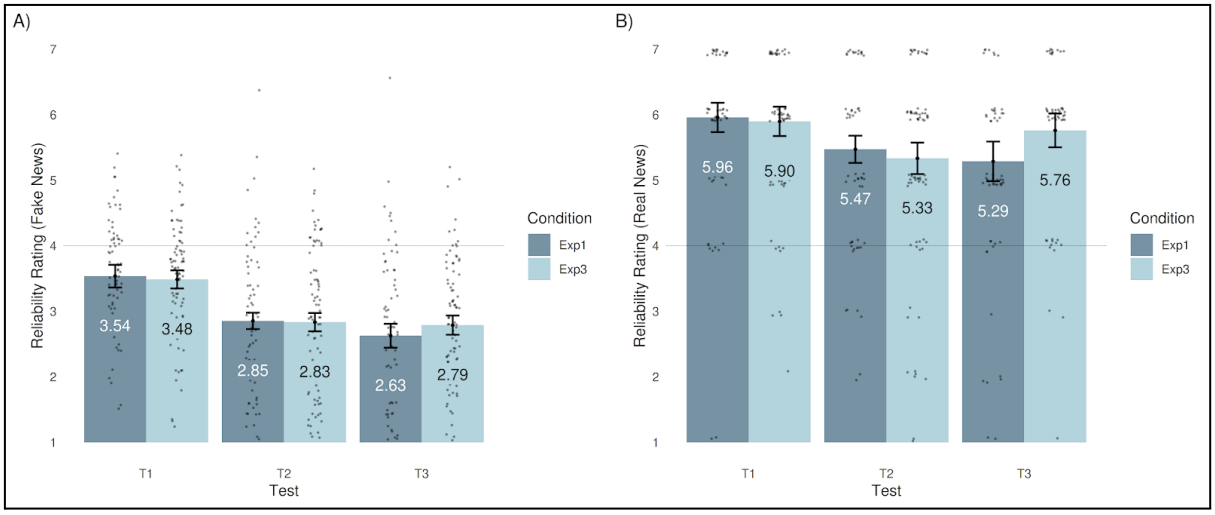

Figure 5.1.6.2. Comparison of reliability ratings of Experiment 3 to Experiment 1. Panel A depicts fake news ratings; Panel B depicts real news ratings. Only items overlapping between both experiments are shown. Horizontal line reflects binary fake (< 4) or real (> 4) classification threshold. Error bars represent 95% confidence intervals. $N = 157$.

Exploratory

Although not pre-registered, we also looked at the real news items. All seven real news items were rated as very reliable (> 4/7) before intervention, immediately after intervention, and one week after intervention. We then compared the two overlapping items that were used in both experiments and found no significant differences between T2-T1 difference scores in Exp 3 ($M_{\text{diff,T2T1}} = -0.56$, $SE = 0.16$) and Exp 1 ($M_{\text{diff,T2T1}} = -0.49$, $SE = 0.13$), with $M_{\text{diff-in-diffs}} = -0.08$, 95% CI_M [-0.48, 0.32], $t(153) = -0.38$, $p = .70$, $d = -0.06$, 95% CI_d [-0.37, 0.25], with statistical equivalence to zero ($t(153) = 2.47$, $p = .007$). This indicated that also for real news, the baseline effect between both experiments was the same.

When comparing the real news T3-T1 difference-in-differences scores between Exp 3 ($M_{diff,T3T1}$ = -0.14, SE = 0.17) and Exp 1 ($M_{diff,T3T1}$ = -0.67, SE = 0.20), we found a significant positive effect with $M_{diff\text{-}in\text{-}diffs}$ = 0.53, 95% CI_M [0.01, 1.05], $t(145) = 2.02$, $p = .045$, $d = 0.33$, 95% CI_d [0.01, 0.64]. A comparable result was found when comparing T3-T2 difference-in-differences scores between Exp 3 ($M_{diff,T3T2}$ = 0.43, SE = 0.18) and Exp 1 ($M_{diff,T3T2}$ = -0.19, SE = 0.19), which showed a significant effect with $M_{diff\text{-}in\text{-}diffs}$ = 0.61, 95% CI_M [0.09, 1.13], $t(151) = 2.31$, $p = .022$, $d = 0.37$, 95% CI_d [0.05, 0.69]. These analyses demonstrated higher reliability ratings for real news at T3 in Exp 3 (where design confounds were removed) as compared to Exp 1 (see Figure 5.1.6.2, Panel B).

Discussion

In Experiment 3, we investigated whether the effects found in Experiment 1 were confounded by the ratio of fake-to-real items, and the repeated use of the same item set. Although we only looked at a time period of one week after the intervention, the results show that there is no significant difference between the results of Experiment 3 and the results of Experiment 1 for fake news. Meanwhile, consistent with Roozenbeek, Maertens, et al. (2021), exploratory analyses indicated that removing the confounds had improved the reliability rating of the real news item.[49] We can, therefore, reasonably conclude that while there may be some longer-term effects of design choices that are not measured here, the findings presented in Experiment 1 and Experiment 2 are unlikely to be due to item-specific or item-ratio effects.

5.1.7 Discussion

Overall, across the three experiments, we successfully replicate the inoculation treatment effect reported by Roozenbeek and van der Linden (2019), but with more rigorous experimental designs. We show that after playing *Bad News*, participants find fake news

[49] It has to be taken into account that this was based on a comparison of one real news item that overlapped between Experiment 3 (T3) and Experiment 1 (T3), and not an index of items as is the case for the fake news analyses.

headlines significantly less reliable than before playing the game. The three inoculation effects (d_{Exp1} = -1.00, d_{Exp2} = -0.69, d_{Exp3} = -0.72) are descriptively larger than the d = -0.52 found in the original study by Roozenbeek and van der Linden (2019). In their broad meta-analysis of inoculation theory, Banas and Rains (2010) found a corrected average of d = 0.43 (95% CI = [.39, .48]) for inoculation interventions compared to control groups over 41 studies. A comparison of these results indicates that the *Bad News* inoculation intervention scores in the high range of inoculation effectiveness. In the broader context of resistance to persuasion research, these can be considered large effect sizes (Weber & Popova, 2012). Given that consequential recent elections have been decided on small margins, practically, these results are also potentially meaningful, especially when applied at population-level (Funder & Ozer, 2019).

Moreover, one potent criticism of such interventions could be that they are potentially less useful if the effects do not persist over time. The field of inoculation research lacks sufficient insights from longitudinal studies in order to accurately draw a decay function of the inoculation effect. Our study provides new insights into the long-term stability of inoculation interventions. Contrary to our expectations, with effects lasting up to at least three months, no evidence was found for the decay of the inoculation effect in Experiment 1. Accordingly, we theorised that regular testing in itself might have a positive "boosting" influence, and thus we leveraged insights from a second experiment (Experiment 2). When we excluded regular follow-ups, the inoculation effect was no longer significant two months after the intervention. These results demonstrate the limits of the longevity of the intervention and add new questions to the debate about the feasibility of long-term resistance against persuasion. The difference between T4 and T5 in Experiment 1 was two months, the same timeframe as between T2 and T3 in Experiment 2. Yet, whereas we find no decay in Experiment 1, we find 64% decay in Experiment 2. The possibility must be considered that in

Experiment 1, little decay was observed because we used the same item sets for each test, meaning that participants may have remembered their responses from the previous test date. Although it is unlikely that they would remember the exact responses one week following the initial test (we had 21 items each on a 7-point Likert scale), the general response tendency could have been remembered. Since the fake-to-real ratio (18:3) was strongly balanced in favour of fake news, this is a valid concern (Aird et al., 2018). We, therefore, conducted a third experiment (Experiment 3), where we presented a different item set for the T3 (one week later) follow-up measure and balanced the fake-to-real item ratio (6:6). Here, we found that the inoculation effect and its decay are not influenced by item memorisation effects, thereby providing a stronger case for a broader "booster shot" or learning mechanism rather than item-specific or simple memory effects. However, these findings cannot fully exclude the possibility that with more follow-up measures response memorisation could play a larger role.

It has been argued that the active inoculation method could be linked to longer retention of the inoculation effect, as, rather than passively reading material, participants are more cognitively involved in the intervention (McGuire, 1961; Rogers & Thistlethwaite, 1969). Researchers have found preliminary evidence that the effect could persist for six weeks up to 33 weeks (Pfau et al., 1992, 2006). In Experiment 1, we find full inoculation retention up to at least 13 weeks, thus pointing towards the potential long-term effectiveness of active inoculation interventions with regular assessment. In Experiment 2, however, we find decay after eight weeks, which may have started within the proposed six-week timeframe for inoculation intervention decay (Ivanov et al., 2018). Future research will have to look deeper into the links between memory strength and inoculation, and the potential of protecting against forgetting by implementing "booster shots." Classical explanations for the decay of the inoculation effect include a decreasing motivation to protect the attacked

attitudes and the lack of a fresh sense of threat (Ivanov, 2017; Miller et al., 2013; Pryor & Steinfatt, 1978). In the context of fake news, we deem it unlikely that the sense of "threat" has disappeared, as fake news has become a common and looming threat in the mainstream media. Moreover, Compton and Ivanov (2012) found that variable testing might boost threat levels and contribute to the effectiveness of inoculation. However, as threat has been shown to be an important contributor to motivation over time (Banas & Richards, 2017) and considering that we did not explicitly measure threat here, we cannot make any conclusions about its role in our study. In addition, linked to threat, Insko (1967, p. 316) stressed that with a decrease in motivation *"the individual ceases to accumulate belief-bolstering material..., [dropping] off over time like the ordinary forgetting curve."* A decreasing motivation is also possible in the context of information overload, as people might start to rely more on heuristics and have less energy to fight against attitudinal attacks (Laato et al., 2020).

However, we argue that an alternative theoretical model could be based on memory strength and forgetting. After an inoculation intervention participants have bolstered their psychological "immune system", but the techniques used in the attacks have to be remembered, and are subject to interference (Hardt et al., 2013). Indeed, we can link various key concepts of inoculation theory to a potential memory model to explain decreases in the inoculation effect. As associative networks have been linked to the long-term memory system (Collins & Loftus, 1975; Smith, 1998), inoculation effect decreases over time could be researched through the lens of neural network simulations of memory networks (Hardt et al., 2013). Over time, the memory network could suffer from *forgetting* (Frankland & Bontempi, 2005), with *interference* as the mechanism (Underwood, 1957).[50] Interference theory refers to forgetting taking place when other (similar or related) information conflicts with (or replaces) the initial memory. "Booster shots" could, therefore, be seen as relearning, protecting against

[50] We do not mention *trace decay* as an explanation as "there has been a long-standing consensus that [trace] decay plays no role in forgetting over the long term" (Brown & Lewandowsky, 2010, p. 51).

interference by strengthening the memory representations (Ebbinghaus, 1885; Ivanov et al., 2018; McGuire, 1961). This leads to the question of whether the decay function can be depicted as a *forgetting curve* (Ebbinghaus, 1885; Murre & Dros, 2015)[51]: an exponential function with the steepness of forgetting being a function of memory strength and time, suggesting that a stronger memory, which can be attained through relearning (cf. booster sessions), will be less susceptible to forgetting (i.e., less influenced by interference).

Accordingly, just like a real vaccine, it might be necessary to have several boosters before long-term immunisation can be established or come to its potential optimum (Compton & Pfau, 2005; Ivanov, 2017). We hypothesise that the tests themselves could have served as "booster shots," being a potential reminder of the techniques and skills learned in the game as well as providing a refreshed sense of threat (Compton and Ivanov, 2012). These findings could also be explained through the lens of memory research, as researchers have shown the importance of repeated testing for memory strengthening (Karpicke & Roediger, 2008; Linton, 1975; Roediger & Karpicke, 2006a, 2006b). A future study could experiment with a shortened or passive version of *Bad News*, for example, to help refresh the cognitive skills participants have acquired during gameplay and to reactivate and strengthen associative memory networks (Pfau et al., 2005).

A different question that remains is how media literacy training can help teach people how to correctly signal real news, as well as fake news. We found that the real news indices used in our study were not reliable. The findings, reported in Supplementary Tables S3-S8 and S12-S13, Supplementary Figures S1-S4, and Figure 5.1.6.2, indicate that in all three experiments real news items remain rated as highly reliable (> 4 out of 7) both before and after intervention, while fake news is rated as low in reliability before and particularly after intervention (< 4 out of 7). In the original large-sample (N = 15,000) study on *Bad News* and

[51] The forgetting curve was proposed by Hermann Ebbinghaus (1885) in his treatise *Über das gedächtnis: Untersuchungen zur experimentellen psychologie* [*On Memory: A contribution to experimental psychology*]. About 130 years later, the forgetting curve was successfully replicated (see Murre & Dros, 2015).

its cross-cultural replication no meaningful change in real news reliability was found, but only two news items were used (Roozenbeek & van der Linden, 2019a). A recent methods paper indicates that negative effects for real news items in *Bad News* may be due to an interaction between the specific item set used and the intervention, and not generalisable to other items (Roozenbeek, Maertens, et al., 2021). Compatible findings were seen in Experiment 3, where the same real news item was rated higher at T3 when the pretest items were different (Experiment 3) than when they were the same (Experiment 1). This finding may be counterintuitive as we know from the illusory truth effect that a repeated presentation of the same headline should be perceived, if anything, as more reliable (Hasher et al., 1977; Hassan & Barber, 2021; Pennycook et al., 2018). One potential explanation for the effect found here is that the inoculation intervention could lead to a general scepticism of all the items seen just before the intervention, while new items after the intervention receive renewed scrutiny. Another potential explanation would be that there is a general scepticism of all news headlines immediately after the intervention, that then disappears in the course of one week. Future research should disentangle testing effects across timepoints, as well as use a more reliable index of real news headlines.

As the measurement in our intervention is the change in reliability that people assign to news messages, we cannot be certain whether any changes in beliefs have occurred. We argue that the reliability rating is a proxy of the readiness to refute the fake item and with this the motivation to protect oneself against it, in line with inoculation theory. Moreover, we caution against the view that news is either “real” or “fake” and that people either “believe” or they “do not”, as most fake news is about subtle degrees of news manipulation (Ecker, Lewandowsky, Chang, et al., 2014; van der Linden & Roozenbeek, 2020). Thus, rather than informing people what is true or false, the *Bad News* intervention trains people to spot misinformation techniques so that people can calibrate their judgments accordingly (Basol et

al., 2020). However, future research measuring shifts in beliefs could help further clarify this distinction.

This study does not come without limitations. The control group, in which people play *Tetris*, does not fully eliminate demand characteristics. In addition, in Experiment 1 and Experiment 2, the item sets used were not balanced in their fake-to-real ratios. Future research could introduce a control group which elicits demand effects and look into the development of a more balanced scale that is equally powerful and reliable for the correct signalling of real news as it is for fake news. Finally, while integrated into the design of the game, like McGuire, we did not explicitly measure threat and motivation. These components have shown to be potentially important in eliciting and maintaining inoculation effects (Banas & Richards, 2017; Compton & Ivanov, 2012), and could provide useful insights into mechanisms behind the longevity of the effect. We recommend future longitudinal studies to explicitly measure these components.

In particular, to unveil the mechanisms of decay, one could consider integrating measures of threat and motivation (Compton & Ivanov, 2012) as it is possible that the treatment effect on fake news ratings is mediated by enhanced threat and motivation (Banas & Richards, 2017; Richards & Banas, 2018). Furthermore, recent best practices suggest the need to square the sample size when testing for interaction effects (Giner-Sorolla, 2018). We recommend that future studies recruit more participants per group, to enable more precise and more generalisable answers about the nature of the decay function. New insights could also be gained by replicating this experiment using more advanced longitudinal designs implementing more time points as well as varying assessment intervals.

In conclusion, with the results of this study, we gain novel insights into the effectiveness and longevity of a real-world fake news intervention based on inoculation theory. In times where the spread of (micro-targeted) misinformation is threatening public

health and scientific and democratic discourse (Lewandowsky et al., 2017), inoculation based interventions could form a crucial part of the solution (Farrell et al., 2019; van der Linden & Roozenbeek, 2020). As the *Bad News* intervention is entertaining, easy to scale, adapt, and tailor, it can be put into action to protect specific groups of people who are most vulnerable to misinformation (Scheufele & Krause, 2019).

Study 4

5.2 The Memory-Motivation Model of Inoculation: Gamified Interventions[52]

5.2.1 Abstract

In this final study on the *Bad News* game, we delved deeper into the questions that remained after Study 3. In a new longitudinal experiment (N = 674), we investigated what the best underlying predictors are related to the retention of inoculation effects over time (T1: 0 days, T2: 9 days, T3: 29 days). More specifically, we tested the memory-motivation model of inoculation, and how predictive it is for the long-term effectiveness of gamified inoculation interventions. In addition, we tested a booster intervention to top up and increase the longevity of the effect. First, we replicate the inoculation main effect at T1. We also found, contrary to expectations, that when no immediate post-test is administered, the effect was no longer significant after 9 days or 29 days. We found that memory was the most dominant factor in all analyses of long-term effectiveness, and that the decay curve follows a pattern that could be reconciled with a forgetting curve. Other predictors, such as motivation and threat, were found to be significant predictors, but do not match the same level of variance that memory explains. Finally, we found that a booster intervention after 9 days significantly increases memory of the intervention at 29 days. This study further corroborated evidence for the memory theory on inoculation, providing evidence for when and why the decay takes place.

[52] Study 4, as depicted here, is a manuscript in preparation as *"The Long-Term Effectiveness of Gamified Inoculation: Mechanisms"*. The paper was written in collaboration with Professor Jon Simons (University of Cambridge), Dr Jon Roozenbeek (University of Cambridge), and Professor Sander van der Linden (University of Cambridge). I am the sole first author on this work.

5.2.2 Introduction

The current study ($N = 674$) is a continuation of Study 3 on the Bad News (BN) paradigm, using a similar methodology and design, with added to it the same additional questions from Study 2 for memory and motivation, and a newly developed version of *Bad News* to serve as the booster game. We set out to shed light on the validity of the memory-motivation theory of inoculation in the setting of gamified inoculation, and to investigate the potential of booster shots further. More specifically, in this study, we sought to test a similar set of hypotheses as in Study 2. We sought to replicate the main effect at T1 (H1) and expected the long-term effectiveness to remain intact for at least 10 days (H2). Meanwhile, we expected the effect to no longer be significant after 30 days when no booster was received (H3), but still significant after 30 days if participants played a booster game 10 days after T1 (H4). We also expected the booster intervention to improve the objective memory of the intervention at T3 (H5), as well as increase the motivation to defend oneself at T3 (H6). Finally, we aimed to test the importance of memory and threat at mediating the inoculation effect (H7).

All preregistered hypotheses are listed in Table 5.2.2.1. A full overview of all items and survey files, R analysis scripts, raw and clean datasets can be found at the OSF repository for this project at https://osf.io/hwmge/?view_only=82bf2bc0f6ec4c5680e728cf5975244a. This study was also preregistered on AsPredicted at https://aspredicted.org/8YF_9L4.

Table 5.2.2.1
Hypotheses of Study 4

#	Hypothesis
H1	People who complete a gamified inoculation intervention (Bad News) rate misleading social media posts as less reliable than people who complete a control task (Tetris).
H2	The inoculation effect described in H1 remains significant for at least 10 days (T2).
H3	The inoculation effect described in H1 is no longer significant after 30 days (T3).
H4	The inoculation effect described in H1 is still significant after 30 days (T3), when individuals participate in a booster intervention after 10 days (T2).
H5	People participating in a booster intervention after 10 days (T2) show increased memory of the inoculation intervention after 30 days (T3) compared to the control group.
H6	People participating in a booster intervention after 10 days (T2) show increased motivational threat after 30 days (T3) compared to the control group.
H7	The inoculation effect [H7a] immediately after intervention (T1), [H7b] after 10 days (T2), and [H7c] after 30 days (T3) is influenced directly by memory and motivation, and indirectly by the inoculation intervention (mediated by memory and motivation).

5.2.3 Methods

Design, Sample, and Procedure

We recruited 1,350 US participants aged 18 or older through Prolific to participate in this study (based on a power = .95, $\alpha = .05$, accounting for up to 50% effect decay). Participants were randomly allocated to an inoculation group with a posttest at T1 only, an inoculation group with a posttest at T2 only (10 days later), an inoculation group with posttest at T3 only (30 days later), the booster group (with posttest at T3 only), or the control group (with posttest at T1, T2, *and* T3). This means that some participants who received an inoculation message at T1 also received a booster treatment at T2 (i.e., those in the booster group). All participants received a pretest at T1 to be used as a covariate during the study. When we refer to T1 in this study, when not otherwise specified, we refer to the posttest at T1. This design was chosen to avoid the boosting by repeated posttesting that we found in Study 3 and enable a clean measure of the long-term effectiveness. We did not separate the groups for the control group as previous studies had shown that the repeated testing effects in the control group were limited (Maertens et al., 2021; Roozenbeek, Maertens, et al., 2021). The time points were chosen to investigate the potential exponential decay between time

points, and as we know from Study 3 that the inoculation effect decays between T1 and 2 months later, and that the literature suggests that decay is likely to be found between 2 weeks (Banas & Rains, 2010; Zerback et al., 2021) and 6 weeks (Ivanov et al., 2018). The specific days between the recruitment were chosen to match the time points used in Study 2. See Table 5.2.3.1 for an overview of the study design.

As preregistered, participants were excluded when they 1) failed the manipulation check (participants were asked to enter a password they received at the end of the intervention), 2) failed both attention checks (e.g., *"The colour test is simple. When asked for your favourite colour you must not select the word puce, but you have to select the word blue. Based on the text you read above, what colour have you been asked to select?"*, [multiple choice, 5 colors]), 3) participated in the survey multiple times, or 4) did not complete the entire survey. We also excluded participants who did not participate in the follow-up within 3 days from the intended participation date.[53] This led to a final sample size of $N = 674$, with an average of 135 participants per group, slightly below the intended $n = 220$ due to a higher-than-expected attrition rate ($T2_{Attrition} = 33.03\%$, $T3_{Attrition} = 47.16\%$). Of the final sample, 54.30% identified as female (41.39% as male; 3.12% as non-binary; 1.04% as transgender, 0.15% as "other"), the average age was 33.18 ($SD = 12.25$, $Mdn = 30$), 53.12% had a higher education degree degree, 66.17% identified as left-wing (22.40% as centrist; 11.42% as right-wing), 68.55% used social media multiple times a day (17.66% once a day, 6.08% weekly, 4.75% less often than weekly, 2.97% never), and 24.63% used Twitter multiple times a day (15.88% once a day, 12.17% weekly, 21.66% less often than weekly, 25.67% never).

[53] It was preregistered that we would exclude participants who did not participate within five days after the intended follow-up date. We chose to change the window to 3 days before or after that date as we sent out grouped invitations manually 1–3 days before the intended follow-up date.

Table 5.2.3.1

Experimental Setup of Study 4

Condition	Time		
	Baseline (T1)	**10 Days (T2)**	**30 Days (T3)**
Control (*n* = 108)	D | **C** | D	D	D
Inoc-T1 (*n* = 211)	D | **BN** | D	-	-
Inoc-T2 (*n* = 134)	D | **BN**	D	-
Inoc-T3 (*n* = 114)	D | **BN**	-	D
Inoc-B-T3 (*n* = 107)	D | **BN**	**B**	D

Note. BN = Bad News game treatment. D = DEPICT reliability rating of 21 news items. C = control task (Tetris). B = booster treatment.

Bad News Booster

While the main intervention uses the same *Bad News* inoculation game as in Study 3, we worked together with the media platform DROG to design a new, shortened, version of the *Bad News* intervention to serve as a "booster treatment". In this 5-minute version of *Bad News*, available at https://www.getbadnews.com/droggame_book/boostershot-bad-news/, participants are asked to put the skills they have learned in the original *Bad News* to use in a new scenario. They have to choose three disinformation techniques they want to revise and then have to use those methods to go through an additional chapter, similar to the original *Bad News*.

Measures

Our main dependent variable was the same variable as used for Study 3, the fake news items of a rating task of 21 social media posts, of which 18 are using one of the six DEPICT disinformation techniques, and 3 real news posts that do not use any of these techniques. Participants were asked to rate each of these news headlines on a reliability rating scale (1–7; M_{Fake} = 2.68, SD_{Fake} = 0.82; M_{Real} = 5.61, SD_{Real} = 0.97). In addition, participants answered the

additional questions from Study 2 on *objective memory* (0–12; $M = 8.97$, $SD = 2.31$), including 4 multiple choice questions (example question: *"What best describes the appealing to emotion misinformation technique?"*; choice options: o *The use of outrage or highly emotive language to manipulate people*, o *Misusing the identity of politicians, experts, or celebrities online*, o *Casting doubt on mainstream narratives by providing an attractive story in which a small sinister group of people is responsible for doing harm to many*, o *Eliciting reactions from people by provoking them online*, o *None of these options is correct*) and 8 yes-or-no questions (overarching question: *Which of the following did you learn about in the game from the first part of the survey?*; example item: *Discrediting your opponent*; response options: o *No*, o *Yes*), *subjective memory* (1–7; $M = 4.36$, $SD = 1.76$), and *interference* (1–7; $M = 3.57$, $SD = 1.76$), as well the questions on *motivational threat* (an index of three Likert items; example item: *Thinking about online misinformation motivates me to resist misinformation*; 1–7, *Strongly disagree–Strongly agree*; $M = 5.19$, $SD = 1.33$), *apprehensive threat* (1–7; $M = 3.49$, $SD = 1.62$), *issue involvement* (1–7; $M = 5.84$, $SD = 2.00$), *post-inoculation talk* (1–7; $M = 2.51$, $SD = 1.39$), *attitude accessibility* (1–7; $M = 3.69$, $SD = 1.55$), and *fear* (1–7; $M = 3.49$, $SD = 1.81$). See the methods section of Study 2 for example questions. When referring to memory in the results section of this study we refer to the outcome of the *objective memory* question, and when referring to motivation we refer to the outcome of the *motivational threat* measure.

5.2.4 Results

Hypothesis Tests

Main Effect

We tested the main effect of the *Bad News* game on participants' reliability rating of misleading content using a one-way ANCOVA with pretest reliability ratings as a covariate, intervention as the independent variable, and misinformation reliability ratings as the

dependent variable, at T1.[54] We found inoculation to have a significant and large effect on the outcome, $F(1, 316) = -43.37$, $p < .001$, $d = -0.779$, 95% CI [-1.020, -0.538], meaning that participants rated fake news as less reliable after the inoculation intervention, and providing strong evidence in favor of H1.

Decay Analysis

The same ANCOVA design as for H1 was used to test the decay hypotheses H2 and H3, this time at T2 (*Mdn* = 9 days after the intervention) and at T3 (*Mdn* = 29 days after the intervention). We found that the inoculation effect was no longer significant 9 days after the intervention, $F(1, 239) = -3.54$, $p = .061$, $d = -0.244$, 95% CI [-0.500, 0.012], contrary to our expectations for H2. At 29 days after the intervention, the omnibus ANCOVA test for the intervention was no longer significant $F(2, 325) = 2.64$, $p = .073$, meaning that neither the single inoculation intervention nor the booster intervention managed to maintain a significant effect, in line with our expectations for H3 but against our expectations for H4. See Figure 5.2.4.1 for an overview of the unreliability ratings (Panel A) and the memory retention (Panel B) over time.

Booster Analysis

As preregistered, we then continued to test whether the booster inoculation had a positive effect on memory of the T1 intervention and motivation at T3. For these analyses we used a T3 ANOVA similar as the one to the previous hypothesis test but this time with memory and motivation as dependent variable for H5 and H6 respectively, and without the pretest. We first found that the used intervention had a significant omnibus effect for memory, $F(2, 326) = 35.56$, $p < .001$, in line with H5, but not for motivation, $F(2, 326) = 0.06$, $p = .966$, leading us to reject H6. Looking at the specific group contrast for memory, we found a

[54] The preregistration proposes a two-way rmANCOVA analysis but the design of this study does not allow us to do this, as participants were separated in different groups for different time-points and the booster group did not receive a posttest before T3. We therefore use a one-way ANCOVA analysis for each time point separately and with pre-test as a covariate instead.

significant and large increase in memory for the booster intervention compared to the control group, $t(326) = 8.43$, $p_{tukey} < .001$, $d = 1.149$, 95% CI [0.867, 1.432], in line with H5. Although not preregistered, we also looked at the difference in memory between the boosted inoculation group and the single inoculation group, also finding a significant difference, $t(326) = 3.99$, $p_{tukey} < .001$, $d = 0.538$, 95% CI [0.270, 0.806]. See Figure 5.2.4.2 for a plot of the effect of memory on unreliability ratings (Panel A) and group on memory (Panel B) for each of the time points.

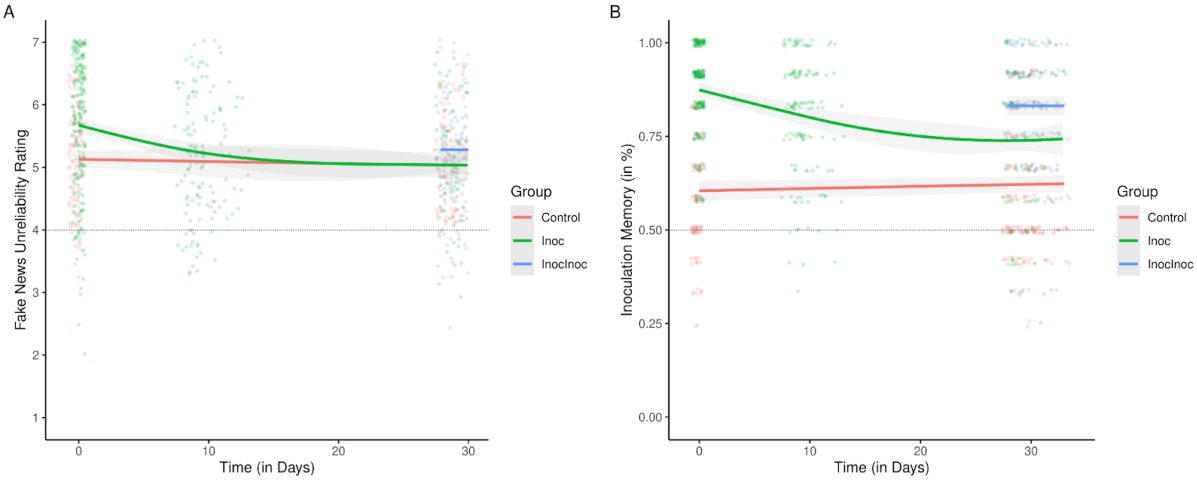

Figure 5.2.4.1. Fake news unreliability ratings (reversed reliability ratings for visualisation purposes) for each time point ($N = 674$). Error bands represent the standard error.

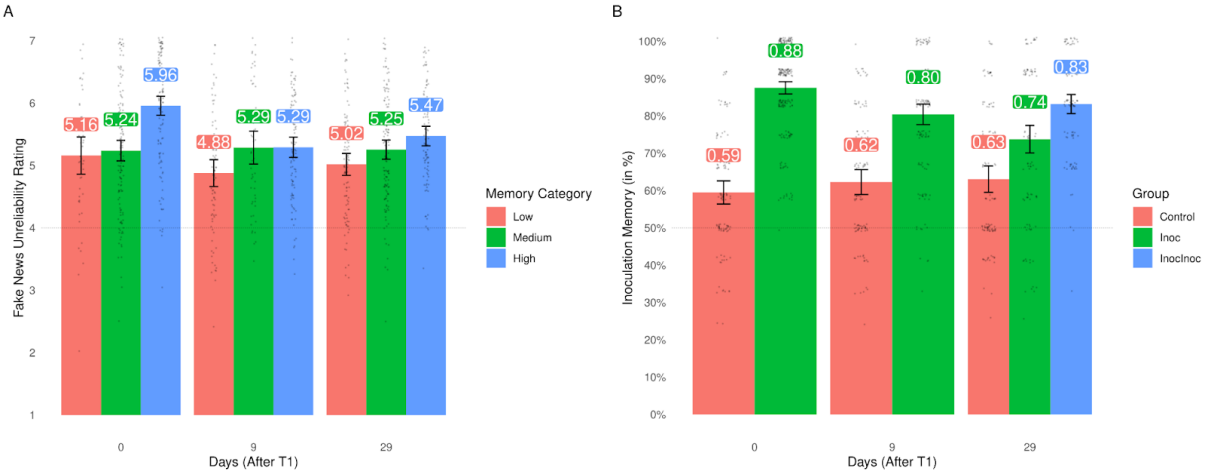

Figure 5.2.4.2. Panel A represents the fake news unreliability ratings for those scoring low, medium, or high on memory in the inoculation group, split by date ($N = 674$). Panel B represents the memory of the inoculation intervention in each group, with a visible memory boost for the group that received a second inoculation after 9 days ($N = 107$). Error bars represent the 95% Confidence Interval.

Memory-Motivation Model

Our final hypothesis is a test of the memory-motivation model of inoculation, to investigate the interplay between memory and motivation in predicting inoculation effect outcome. To do this, as preregistered, we tested an SEM model using *lavaan* in R (Rosseel, 2012) that includes inoculation as a predictor of the fake news detection score, and memory and motivation as mediators. We found, in line with [H7], that memory had a direct influence on fake news reliability ratings at T1 [H7a], $z = -5.10$, $p < .001$, $\beta = -0.372$, 95% CI [-0.515, -0.229], at T2 [H7b], $z = -3.14$, $p = .002$, $\beta = -0.225$, 95% CI [-0.365, -0.084], and at T3 [H7c], $z = -4.16$, $p < .001$, $\beta = -0.242$, 95% CI [-0.355, -0.128]. However, motivational threat was not a significant predictor of fake news reliability ratings at T1 [H7a], $z = -1.85$, $p = .064$, $\beta = -0.097$, 95% CI [-0.199, 0.006], at T2 [H7b], $z = -0.15$, $p = .883$, $\beta = -0.009$, 95% CI [-0.133, 0.114], or at T3 [H7c], $z = -1.18$, $p = .238$, $\beta = -0.064$, 95% CI [-0.169, 0.042]. Motivation did significantly influence memory formation at T1, $z = 2.13$, $p = .033$, $\beta = 0.085$, 95% CI [0.007, 0.163], in line with the memory-motivation model.

Further in line with the memory hypothesis of H7, inoculation had an indirect effect on fake news detection outcome mediated through memory at T1 [H7a], $z = -4.90$, $p < .001$, $\beta = -0.548$, 95% CI [-0.767, -0.329], at T2 [H7b], $z = -2.94$, $p = .003$, $\beta = -0.215$, 95% CI [-0.358, -0.072], and at T3 [H7c], $z = -3.62$, $p < .001$, $\beta = -0.192$, 95% CI [-0.296, -0.088]. Although not preregistered, we also looked at whether the direct effect of the inoculation intervention was still significant at T1 when accounting for memory, and we found that the direct effect was no longer significant, $z = 0.25$, $p = .803$, $\beta = 0.038$, 95% CI [-0.262, 0.338].

See Figure 5.2.4.3 for a schematic presentation of the tested T1 approximation of the memory-motivation model and Table 5.2.4.4 for its model estimates.

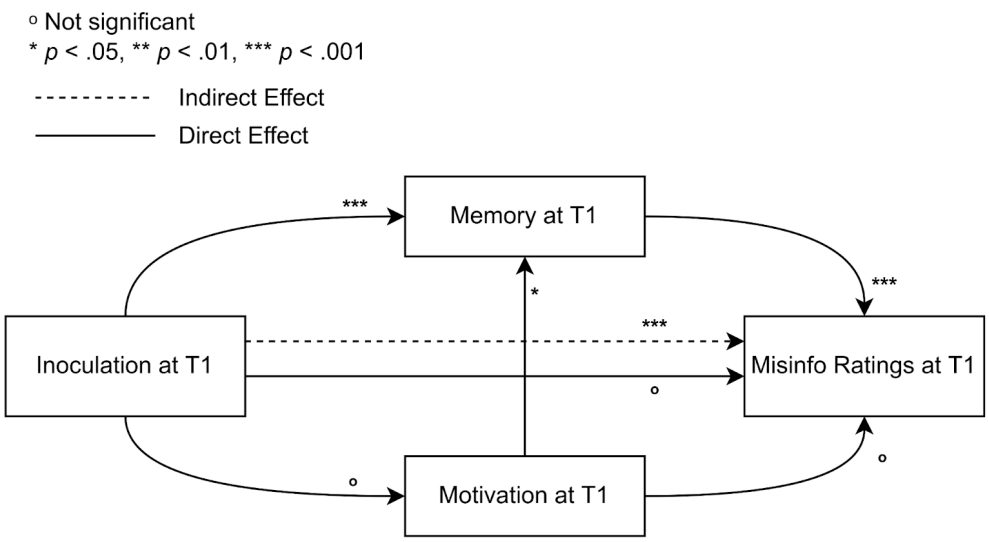

Figure 5.2.4.3. SEM analysis of the memory-motivation model at T1 in Study 4 (*N* = 319).

Table 5.2.4.4

Memory-Motivation Model Estimates at T1 in Study 4, N = *319*

Effect	*z*	*p*	β	95% CI		*SE*
				LL	*UL*	
Indirect						
Inoc.T1 ⇒ Memory.T1 ⇒ Fake.T1	**-4.898**	**< .001**	**-0.548**	**-0.767**	**-0.329**	**0.112**
Inoc.T1 ⇒ Motivation.T1 ⇒ Fake.T1	-0.452	.651	-0.005	-0.028	0.018	0.012
Inoc.T1 ⇒ Motivation.T1 ⇒ Memory.T1 ⇒ Fake.T1	-0.454	.650	-0.002	-0.009	0.006	0.004
Component						
Inoc.T1 ⇒ Memory.T1	**17.564**	**< .001**	**1.472**	**1.308**	**1.636**	**0.084**
Memory.T1 ⇒ Fake.T1	**-5.100**	**< .001**	**-0.372**	**-0.515**	**-0.229**	**0.073**
Inoc.T1 ⇒ Motivation.T1	0.466	.641	0.055	-0.176	0.287	0.118
Motivation.T1 ⇒ Fake.T1	-1.853	.064	-0.097	-0.199	0.006	0.052
Motivation.T1 ⇒ Memory.T1	**2.133**	**.033**	**0.085**	**0.007**	**0.163**	**0.040**
Direct						
Inoc.T1 ⇒ Fake.T1	0.249	.803	0.038	-0.262	0.338	0.153
Total						
Inoc.T1 ⇒ Fake.T1	**-4.503**	**< .001**	**-0.517**	**-0.741**	**-0.292**	**0.115**

Exploratory Analyses

Dominance Analysis

We performed a dominance analysis on the possible predictors of the fake news reliability rating at T3 (see the methods section of Study 2 for an explanation of dominance

analysis). We found that memory was the dominant predictor, followed by motivational threat. See Table 5.2.4.5 for an overview of the dominance analysis outcome. In addition, although not preregistered, a Pearson correlation test reveals a significant negative correlation between memory and fake news reliability ratings in the inoculated groups, $t(564) = -8.69$, $p < .001$, $r = -.344$, 95% CI [-.414, -.269], as well as a significant negative correlation between memory and time, $t(564) = -5.77$, $p < .001$, $r = -.236$, 95% CI [-.312, -.157], similar to the positive correlation between fake news reliability ratings in the inoculation group and time, $t(564) = 3.94$, $p < .001$, $r = .164$, 95% CI [.082, .243]. A visual analysis of the association between inoculation intervention memory and fake news reliability ratings can be found in Figure 5.2.4.6.

Table 5.2.4.5

Dominance Analysis in Study 4, at T3, N = *329*

Variable	Dominance
Memory	60%
Motivational Threat	17%
Issue Talk	14%
Issue Accessibility	7%
Self-Reported Remembrance	1%
Issue Involvement	1%
Apprehensive Threat	1%
Fear	0%

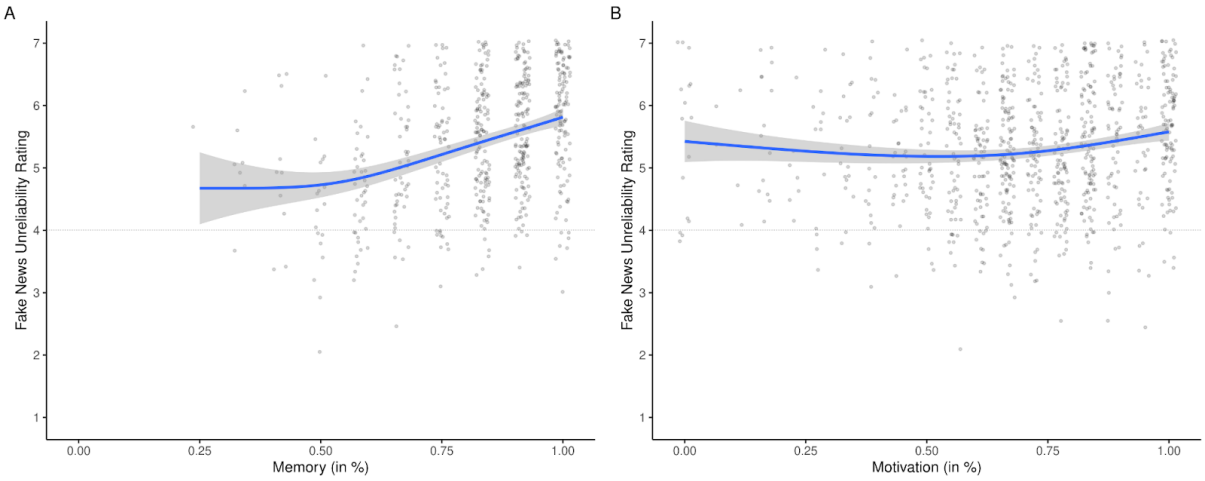

Figure 5.2.4.6. Panel A: the correlation between memory and the unreliability rating of fake news items, for those who received the inoculation intervention (N = 566). Panel B: the correlation between motivational threat and the unreliability rating of fake news items for all participants (N = 674). Error bands represent the standard error.

5.2.5 Discussion

In Study 4 we investigated the new questions raised in Study 3, and explored the feasibility of a memory-motivation model of inoculation within a gamified inoculation paradigm. In the experiment, we first replicated the large main effect at T1 (d = -0.779), but unexpectedly, we found that the effect was no longer significant 9 days after the intervention and therefore also not at 29 days. These findings are more in line with the findings by Zerback et al. (2021), who found a full inoculation effect decay after two weeks, than the findings by meta-analysis by Banas and Rains (2010), who found that typically inoculation effects remain significant for two weeks. Also the booster intervention, although descriptively improving the fake news reliability ratings at 29 days, was not sufficient to uphold the effects for a month. Exploring the underlying mechanisms revealed the traditional variables linked to inoculation effects, *motivational* and *apprehensive threat* (Banas & Richards, 2017; Richards & Banas, 2018), *post-inoculation talk* (i.e., do you talk about the

inoculation content), *fear* (i.e., being fearful about misinformation threats), *issue involvement* (i.e., how much do you engage with the topic), and *attitude accessibility* (i.e., how often do you think about your own attitude towards the topic; Dillingham & Ivanov, 2016; Ivanov et al., 2012; Pfau et al., 2003, 2004, 2005), showed that they only explain a limited amount of variance as compared to memory of the inoculation game's content (memory = 60%, motivation = 17%), with strong dominance of memory over all other predictors.

An SEM test of the memory-motivation model showed significant indirect effects of the inoculation effect through memory, but no significant effects of motivation on the intervention outcome, nor did the inoculation game influence people's motivation. The only significant effect found for motivation was on memory formation at T1, which might indicate that motivation is especially important for the formation of the initial memory. Moreover, we found a significant negative correlation between memory and fake news reliability ratings, meaning that those with a better memory of the inoculation intervention rated misinformation as less reliable. In line with what we found for the climate change inoculation paradigm in Study 2, we found that only the indirect effect of the inoculation intervention was significant (mediated by memory), and the direct effect was not. Although the SEM modelling allowed us to test direct and indirect effects of the inoculation intervention with memory and motivation as mediators, there are three limitations. As we recruited a separate sample for each follow-up date, we were not able to relate T1 variables with variables at 9 days (T2) or at 29 days (T3) using direct paths. Another issue is that the current design does not allow us to fully disentangle the causal order of the model, as memory and motivation were measured at the same time and not manipulated separately, it is also possible that memory increased motivation instead of the other way around. Finally, the intervention was not successful at increasing motivation, which may mean it does not utilize all of the mechanisms that are expected to be found in an inoculation intervention. Future research could manipulate

memory and motivation separately, and consider having a condition that enables the direct path modelling of the variables between different time points.

The lack of significant inoculation effects after 9 days and after 29 days for the booster intervention group could be in part due to the lack of an immediate posttest after the interventions. In previous experiments with *Bad News* (see e.g., Roozenbeek & van der Linden, 2019a; Maertens et al., 2021), an immediate posttest was always presented to the participants, where they had to use the insights of the intervention applied to the rating of the reliability of news items. This could have helped with learning and with the translation of the materials of the intervention into the skill of misinformation detection, as well as a stronger memory foundation. Although research indicates that a pretest does not change the intervention outcome of fake news ratings in *Bad News* studies (Roozenbeek, Maertens, et al., 2021), future research should try to disentangle the effect of an immediate posttest on memory formation and the long-term effectiveness of inoculation interventions. A related limitation with the current study is that we only had sufficiently reliable items for the ratings of misinformation, but not for the ratings of real news, and that the news items did not have a balanced ratio of fake news to real news. Although ratio effects have been discussed by Maertens et al. (2021), it could be argued that this provides an incomplete perspective of the intervention's effects. We recommend that future studies invest equally in the real news headlines to be able to compare them validly, even if the focus of the intervention is just on fake news.

Whether the booster intervention was successful or not is uncertain based on the current findings. Although the effect was no longer significant after 29 days even for the boosted group, it may have decayed as well, as we do not know what the effect was immediately after the booster. However, we do know that the booster had a significant effect on memory after 29 days compared to the non-boosted inoculation group ($d = 0.538$), and

therefore we can say that the booster had an effect. The results also reveal that the inoculation memory decays over time in a similar fashion as the inoculation effect, and both decay in a way similar to an exponential forgetting curve (Ebbinghaus, 1885; Murre & Dros, 2015).

Taken together, these findings provide further evidence for the memory-motivation model's validity, and specifically, for its compatibility with gamified inoculation interventions, in a similar way as was found for the climate change paradigm in Study 2.

Chapter 6

Video-Based Inoculation

In this chapter, I explore the memory-motivation theory assumptions with the video-based inoculation paradigm, which is the newest and least explored type of inoculation. This project comes forth from a collaboration with Google, who provided a grant to design and test new inoculation videos that can be employed at scale as educational advertisements. In a three-experiment study (**Study 5**), I explore the effectiveness of long and short inoculation videos (**Study 5, Experiment 1**), map the long-term effectiveness (**Study 5, Experiment 2**), and investigate the mechanisms behind the effects (**Study 5, Experiment 3**). Similar to what was found in **Chapter 5**, I find that an immediate post-test increases the longevity of the intervention, and that a short inoculation booster video can strengthen memory. Meanwhile, when an immediate posttest is used, the effectiveness of the intervention can last up to at least 29 days without any other booster intervention. I also replicate the finding that memory is the most dominant predictor of the effect longevity, but that motivation is important as well, in line with the other two inoculation paradigms, and provide a third evidence base for the memory-motivation model of inoculation. All materials, clean and raw datasets, survey files, and analysis scripts can be found on the OSF repository for this study at https://osf.io/zrq87/?view_only=375c0632fca0444fa07c2bc46a59187b.

Study 5

6.1 The Long-Term Effectiveness of Inoculation Against Misinformation & The Memory-Motivation Model of Inoculation: Video-Based Interventions[55]

6.1.1 Abstract

News consumption increasingly takes place on social media and video platforms such as YouTube, platforms where misinformation proliferates. Inoculation theory provides a useful framework to protect people against misinformation, however, interventions have been difficult to scale, and insights on their long-term effectiveness have been lacking. In this work we develop and test new, short, video-based inoculation interventions over time and unveil the underlying mechanisms in three pre-registered longitudinal experiments with large US representative quota samples. In Experiment 1 (N = 2,219), we compared a long inoculation video (1 minute 48 seconds) to a short inoculation video (30 seconds), and found that both are similarly effective at improving discernment of misleading content that utilizes emotional language compared to a long (1 minute 46 seconds) or a short (30 seconds) control video. Then, in Experiment 2 (N = 4,821), we longitudinally track the effectiveness of the short inoculation video over time, and explore the underlying mechanisms responsible for its effectiveness. Finally, in Experiment 3 (N = 2,220), we investigate the potential of a "booster" video to maintain the long-term effectiveness of the intervention. We find that the inoculation effect remains intact for at least one month when an immediate posttest is used, but that the effect decays over time. We also find that the "booster" video is effective at protecting the effects from decay. We shed light on the driving mechanisms behind these

[55] Study 5, as depicted here, is a manuscript in preparation as *"The Long-Term Effectiveness of Video-Based Inoculation: Three Longitudinal Experiments"*. The paper was written in collaboration with Dr Jon Roozenbeek (University of Cambridge), Professor Stephan Lewandowsky (University of Bristol), Vanessa Maturo (Google Jigsaw), Rachel Xu (Google Jigsaw), Beth Goldberg (Google Jigsaw), and Professor Sander van der Linden (University of Cambridge). I am the sole first author on this work.

effects, showing that memory of the initial intervention is more important than the perceived threat from misinformation and the motivation to protect oneself against it. We conclude with concrete suggestions for practitioners on how to implement inoculation videos at a large scale while preserving the benefits.

6.1.2 Introduction

In modern society, misinformation proliferates widely and mutates in many different forms (Lewandowsky, 2020; van der Linden, 2022), including in video format (Donzelli et al., 2018; Hussein et al., 2020; Li et al., 2020). Although most people do not consume large quantities of outright fake news (Allen et al., 2020; Grinberg et al., 2019; Nelson & Taneja, 2018), a person can be mislead by subtle manipulation using techniques such as appealing to emotion (Carrasco-Farré et al., 2022; Roozenbeek, Traberg, et al., 2022), and misinformation is known to have a tangible negative impact on society. Moreover, misleading information has been linked to radicalization and extremism (Garry et al., 2021; Zihiri et al., 2022), terrorist propaganda (Chiluwa, 2019; Piazza, 2022), and vaccine hesitancy (Loomba et al., 2021; Pierri et al., 2022).

Due to the fast and far spread of misinformation (Vosoughi et al., 2018), recent research has explored the most effective methods to counter misinformation. One effective method is debunking, a reactive method which tries to undo the damage done by misinformation by correcting false claims (Chan et al., 2017). Another method is nudging people to pause and think or to highlight accuracy while they are making their judgment about a news item (Fazio, 2020; Pennycook et al., 2020). However, both methods are not sufficient to prevent the influence of the exceedingly wide range of misinformation in circulation. A third, more proactive method, can be found in inoculation theory (McGuire, 1961; McGuire & Papageorgis, 1961).

Inoculation Theory

For over 60 years, the most important framework of resistance against persuasion has been inoculation theory (Traberg et al., 2022). An inoculation intervention combines a cognitive and an affective component to confer resistance against an upcoming attitudinal threat (Ivanov & Parrott, 2017). The cognitive component, also known as the preemptive refutation, teaches people about the misleading content and its flaws, often by exposing them to a weakened version of the argument (e.g., a claim with the reasoning flaws highlighted), and how they can resist it. The affective component, often implemented as a forewarning, helps to make people aware of the threat of misinformation and that they are vulnerable and could be exposed to an attitudinal attack, and thereby motivates them to defend themselves and counterargue (Compton, 2013). In addition, the inoculation intervention can be focused, tackling a specific misinformation message, or broad, focusing on a general type of misleadingness (Roozenbeek & van der Linden, 2019a, 2019b; van der Linden & Roozenbeek, 2020). However, as these interventions are typically difficult to scale—the currently best-known gamified inoculation intervention managed to reach more than 2 million players in 3 years (Roozenbeek & van der Linden, 2019a), but this is an exception and one could argue still not enough—we need to consider new avenues where inoculation has a higher potential to spread wide and fast.

Video-Based Inoculation

While inoculation interventions were initially explored with trivial experiments on conferring resistance against persuasive challenges about toothbrushing using essays (McGuire, 1970), it recently took on more polarized issues such as climate change misinformation (Cook et al., 2017; Maertens et al., 2020), and new formats such as gamified interventions (Cook et al., 2022; Roozenbeek & van der Linden, 2019a). Due to the increasing role of videos in the news consumption diet—according to Pew Research Center

(2020, 2021), 81% of Americans use YouTube, and 26% use it for news consumption—researchers have also started exploring the potential of media literacy videos, with promising first results (Lim & Ki, 2007; Nabi, 2003; Pfau et al., 1992, 2000; Varker & Devilly, 2012; Vraga et al., 2021). However, much about the underlying mechanisms of video-based inoculation, and how it compares to other inoculation interventions, remains unknown.

Longevity

Despite the long history of inoculation and successful development of novel interventions, much remains unknown about the mechanisms driving the long-term effectiveness (or the decay) of the inoculation effects (Banas & Rains, 2010; Traberg et al., 2022). While recent studies have shown that the effect of inoculation can become insignificant after two weeks (Zerback et al., 2021) or six weeks (Ivanov et al., 2018), or remain fully effective when "booster" interventions are used (Maertens et al., 2021), the actual decay curve and the underlying mechanisms have not been revealed (Banas & Rains, 2010; Maertens et al., 2021; Ivanov et al., 2018). Two mechanisms that have been proposed to drive both the main effect of inoculation and its longevity, namely the motivation to defend yourself against misinformation (motivational threat; Banas & Richards, 2017) and memory of the inoculation intervention (Maertens et al., 2021; Pfau et al., 2005). However, these two variables had never been systematically explored together and in comparison to each other in a longitudinal experimental study, other than in Study 2 and Study 4 presented in this thesis.

The Present Study

In the current study we set out to explore the long-term effectiveness of short, video-based inoculation interventions, as well as the mechanisms driving these effects, in three longitudinal experiments. In particular, with our first experiment and set of hypotheses, we explored the effectiveness of a short compared to a long inoculation video (see Table

6.1.2.1, Experiment 1). In our second experiment and set of hypotheses, we tentatively explored the role of memory and motivational threat, as well as the longevity of the inoculation effect using multiple time points (see Table 6.1.2.1, Experiment 2). In our third and final experiment and set of hypotheses see Table 6.1.2.1, Experiment 3), we attempted to replicate the findings from Experiment 2, test the memory-motivation model from Chapter 2 with a set-up comparable to Study 2 and Study 4, and explore the potential role of three types of "booster" interventions (a threat-focused video, an inoculation memory rehearsal video, and repetition of the original inoculation video) to disentangle the most effective method to boost inoculation effects in video-based interventions.

Table 6.1.2.1

Preregistered Hypotheses of Study 5

#	Hypothesis
	EXPERIMENT 1
H1.1	People who watched a short inoculation video (H1.1a) or a long inoculation video (H1.1b) are better at discerning manipulative social media posts from neutral social media posts than people who watched a control video.
H1.2	The inoculation effect of a short inoculation video (0 min 30 sec) is smaller than the inoculation effect of a long inoculation video (1 min 48 sec).
H1.3	The inoculation effect of a long inoculation video (H1.3a) or a short inoculation video (H1.3b) decays partially but not completely over a period of two weeks.
	EXPERIMENT 2
H2.1	People who watched an inoculation video are better at discerning manipulative social media posts from neutral social media posts than people who watched a control video.
H2.2	There is no decay of the inoculation effect of inoculation videos after 4 days.
H2.3	There is partial decay of the inoculation effect of inoculation videos after 10 days.
H2.4	There is full decay of the inoculation effect of inoculation videos after 30 days.
H2.5	Memory (forgetting) predicts inoculation decay.
H2.6	Threat (motivation) does not predict inoculation decay.
	EXPERIMENT 3
H3.1	People who watched an inoculation video are better at discerning manipulative social media posts from neutral social media posts than people who watched a control video.
H3.2	The inoculation effect of inoculation videos is no longer significant after 30 days.
H3.3	An inoculation video (T1) that is followed by a threat-based booster video 10 days later (T2), is effective at keeping the inoculation effect significant up to 30 days after the T1 inoculation.

H3.4	An inoculation video (T1) that is followed by a memory-based booster video 10 days later (T2), is effective at keeping the inoculation effect significant up to 30 days after the T1 inoculation.
H3.5	An inoculation video (T1) that is followed by the same inoculation video 10 days later (T2), is effective at keeping the inoculation effect significant up to 30 days after the T1 inoculation.
H3.6	Groups exposed to a threat-based booster video at T2 show increased motivation (a), but not memory (b) of the intervention, at T3, compared to those inoculated who did not receive a booster video.
H3.7	Groups exposed to a memory-based booster video at T2 show increased memory (a) of the inoculation intervention, but not motivation (b), at T3, compared to those inoculated who did not receive a booster video.
H3.8	Groups exposed to a repeated-inoculation booster video at T2 show increased memory (a) of the inoculation intervention and motivation (b) at T3, compared to those inoculated who did not receive a booster video.
H3.9	The inoculation effect at T1 (a) and T3 (b) is influenced directly by memory and motivation, and indirectly by the inoculation intervention (mediated by memory and motivation).

6.1.3 Methods

Procedure

In all three experiments of this study participants were recruited and rewarded for their participation by *Respondi* (an ISO-certified online panel provider). All samples were representative quota samples of the United States based on the age and gender composition data provided by the United States Census Bureau (2019). After recruitment and informed consent, participants took part in a Qualtrics survey and were randomly allocated to one specific condition, followed by a posttest, and in some cases a follow-up. The study was approved by the Cambridge University Psychology Research Ethics Committee (ref. PRE.2021.012). All datasets, analysis scripts in R, Qualtrics surveys, preregistrations, and stimuli are available on the OSF repository at https://osf.io/zrq87/?view_only=375c0632fca0444fa07c2bc46a59187b.

Intervention Videos

For this study we built on a previous study where the researchers tested a range of new inoculation videos (Roozenbeek, van der Linden, et al., 2022). One of the most effective

videos was a video inoculating against manipulative content using emotional language. There is a wide literature showing the importance of emotion in believing and sharing misleading news, and that it is being used as a technique to influence or manipulate people (Bakir & McStay, 2018; Brady et al., 2017; Carrasco-Farré, 2022; Martel et al, 2020; Van Bavel et al., 2021; Vosoughi et al., 2018). Further building on this work, we collaborated with *Studio You London* to create a set of new inoculation videos based on the previous study. More specifically, we created a short version (30 seconds) of the original inoculation video (1 minute 48 seconds), a short booster video focusing on strengthening memory of the inoculation intervention (30 seconds), and a short booster video to reinvigorate a sense of threat (26 seconds). The control videos were selected to be relevant, informative, and of a similar length to the original video. They were taken from YouTube and Vimeo, and were about freezer burn (1 minute 46 seconds) and age-related macular degeneration (30 seconds) for the long and short video respectively.

For the development of each video, we first wrote a script with the research team that contained the two essential elements in an inoculation treatment: a warning message and an explanation of the manipulative technique. Then we discussed and updated the scripts together with the film studio, followed by a back-and-forth of video drafts, leading to a final version of the video. In the video scripts, we made sure that both the long inoculation video and the short inoculation video contained a warning message as well as a cognitive training element to recognise the underlying manipulative technique (emotional language). With the warning message we wanted to maximise the sense of threat a participant perceives, and used a sad child (both for the long inoculation and the short inoculation videos) or an elephant (Booster A) combined with emotional music and narration (e.g., *"You might be thinking about skipping this ad…. Don't. What happens next will make you tear up."*) to attract people's attention. After that, the watchers are told they got tricked into watching this video

by appealing to their emotion. This is then followed by an explanation of the underlying techniques (e.g., *"appealing to emotions like fear or outrage is a trick to get you to pay attention and is key for the spread of misleading content through social networks"*), and additional examples (i.e., a "microdose" of fake news) where participants are asked to think about how this can be applied to make headlines more manipulative (e.g., "*if a ruling is 'disagreeable', call it 'disgusting'"*). In the threat booster video (Booster A) the animated content and dialogue was slightly different as to make sure the activation of memory was minimised and did not include an explanation of the techniques, but included a similar warning part where the participants are lured into watching for longer after seeing an elephant with sad background music. In the memory booster video (Booster B), the sad music is replaced with upbeat music and the warning message is stripped away (only the sad child remains), but the techniques are explained in detail with the same materials as in the original video. A screenshot of the memory-boosting inoculation video can be found in Figure 6.1.3.1. All inoculation and control videos, as well as the complete video scripts, can be found on the OSF repository at https://osf.io/zrq87/?view_only=375c0632fca0444fa07c2bc46a59187b.

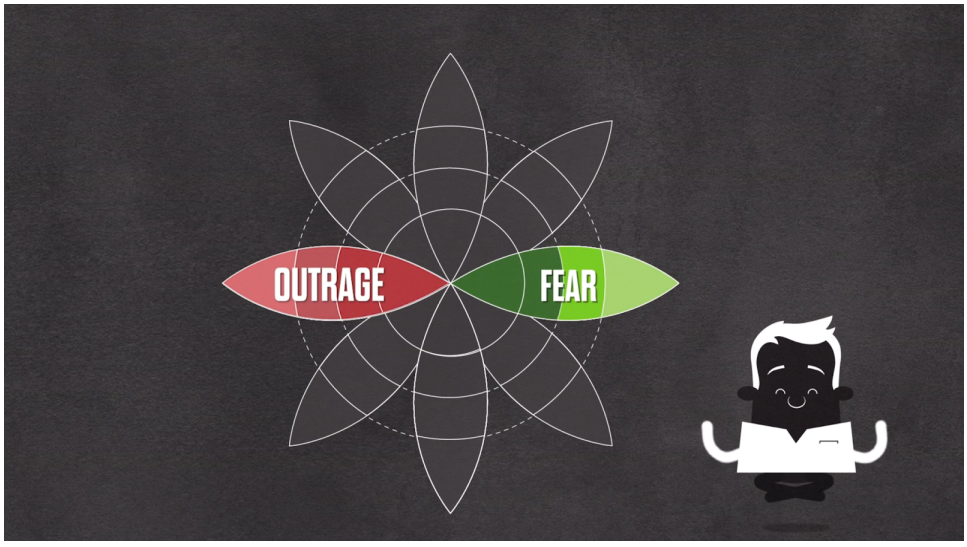

Figure 6.1.3.1. Screenshot of the memory-boosting inoculation video.

Measures

After watching a video, participants completed a social media post rating task, which involved rating a series of ten either manipulative (i.e., containing a manipulation technique) or neutral (i.e., not using any manipulation) social media posts that were based on actual news in the field. The 10 headlines participants rated came from a pool of 20 items consisting of 10 pairs: for each news story we created a manipulative version and a non-manipulative version conveying the same message, and participants were randomly allocated a neutral or manipulative version of each pair, and all social cues (e.g., likes, names, sources) were redacted from the items. This also meant that the manipulative-to-neutral item ratio varied among participants. This setup allowed us to calculate a clean discernment index without the influence of topics, social cues, or item ratios. Specifically, participants were asked to indicate for each post 1) how manipulative they found the post (our main dependent variable for this study); 2) how confident they were in their ability to assess the post's manipulativeness; 3) how trustworthy they found the post; and 4) how likely they were to share the post with others in their network. This rating task was our main method of assessing the videos' efficacy in terms of improving participants' ability to identify manipulative content: if the inoculation videos are effective, treatment group participants should be significantly better than a control group at discerning manipulative from non-manipulative content, have significantly higher confidence in their ability to do so, find manipulative content less trustworthy than neutral content, and should display significantly less sharing intentions for manipulative content than for neutral content. See Figure 6.1.3.2 for an example item.

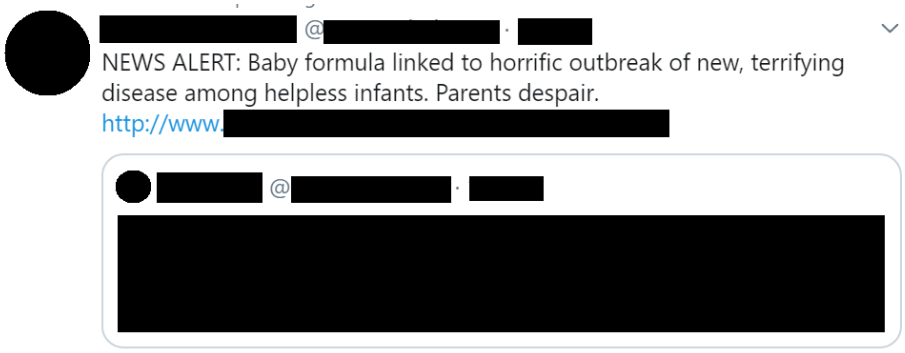

Figure 6.1.3.2. Example of a misleading social media item used in Study 5.

In addition, we investigated the underlying mechanisms of the inoculation effect in line with Study 2 and Study 4. We asked a set of questions to assess participants' sense of threat about emotional language on social media and related constructs (adapted from the measures used by Banas & Richards, 2017; Dillingham & Ivanov, 2016; Ivanov et al., 2012; Pfau et al., 2003, 2004, 2005; Richards & Banas, 2018). with measures for *apprehensive threat* (1–7; $M = 3.32$, $SD = 1.67$), *fear* (1–7; $M = 3.00$, $SD = 1.77$), *issue involvement* (1–7; $M = 4.81$, $SD = 2.49$), *attitudinal accessibility* (1–7; $M = 4.81$, $SD = 2.49$), talk about the issue (*issue talk*; 1–7; $M = 2.37$, $SD = 1.38$), their motivation to counter-argue against emotional headlines (*motivational threat*; an index of 3 Likert items; example item: *"Thinking about the idea of emotional language on social media motivates me to resist misinformation"*, 1 *strongly disagree*, to 7 *strongly agree*; $M = 4.88$, $SD = 1.45$). We also asked a series of questions designed to assess how well people remembered the lessons from the inoculation video, including *self-reported remembrance* (1–7; $M = 3.61$, $SD = 1.90$), *interference* (1–7; 2.76, $SD = 1.61$), and an objective memory test (0–12; $M = 7.31$, $SD = 2.43$) consisting of 4 multiple choice questions (example item: *What example was given in the video for "using emotional language in news headlines"?*; choice options: o *Changing a headline from "serious accident" into "horrific accident"*, o *Using a radio broadcast to trigger emotions*, o *Triggering emotions by employing emojis*, o *None of these options is*

correct) and 8 yes-or-no questions (general question: *Which of the following did you learn about in the video that you watched in part 1 of this survey?*; example entry: *The role of fear and outrage*; choice options: o *No*, o *Yes*). See Chapter 2 and Chapter 4, Study 2 (Methods), for a more detailed discussion of these measures. As an exploratory measure we also created a measure for *concept mapping*, in which participants had to write down as many concepts related to the theme and intervention as possible in open boxes (*Please write down as many concepts or ideas you learned from the video in Part 1 of the survey as you remember*; 0–9; $M = 1.90$, $SD = 1.73$), inspired by the memory concept mapping method by Pfau et al. (2005).

To explore further covariates we also measured *misinformation susceptibility* (as measured through the 8-item Misinformation Susceptibility Test or MIST-8; Maertens et al., 2022; 0–8; $M = 5.98$, $SD = 1.70$), *conspiracy mentality* (CMQ; Bruder et al., 2013; 1–7; $M = 4.58$, $SD = 1.30$), the level of *trust* in politicians, family members, journalists, and civil servants, *party affiliation*, *political self-identification*, and self-reported *ideology* in terms of social (1–7; $M = 3.96$, $SD = 1.74$) and economic (1–7; $M = 4.35$, $SD = 1.70$) issues. Finally, all participants responded to the same series of demographic questions: age, gender, education level, racial background, country of residence, news consumption behavior, whether English is their first language, and their favorite media outlet.

The Qualtrics files and the full PDF printout of the surveys can be found on the OSF repository for this study at https://osf.io/zrq87/?view_only=375c0632fca0444fa07c2bc46a59187b.

6.1.4 Experiment 1

Methods

Sample and Design

The goals of the first experiment were as follows: 1) to replicate the original inoculation video study's findings (Roozenbeek, van der Linden, et al., 2022), 2) to identify

differential effect sizes depending on video length (the full-length 1:48 min video and its shorter version of 0:30 min), 3) to determine the decay percentage after a two-week period, 4) to explore the role of memory and threat in inoculation effects, and 5) to explore if the inoculation effect is moderated by covariates such as conspiratorial thinking, misinformation susceptibility, and political polarization. To answer these questions, we conducted a preregistered longitudinal randomized controlled trial by using the ISO-certified online panel provider *Respondi* with a nationally representative age and gender based quota sample of the US population with power = 0.95 and alpha = 0.05 for an effect size of $d = 0.490$ (based on Roozenbeek, van der Linden, et al., 2022). The recruited sample size was $N = 2{,}895$, with a reduction to $N = 2{,}219$ when counting complete responses only and after—as preregistered—removing participants who failed both the manipulation check (*"Which video did you watch earlier in this study?"*, [multiple choice, 8 descriptions]) and the attention check (*"Please do not select "blue", but select "purple""*, [multiple choice, 9 colors]), participated multiple times, or entered the same response to each of the items of the dependent variable. In our final sample, 50.70% identified as female (48.26% as male; 0.86% as non-binary; 0.05% as "other"; 0.14% preferred not to answer), the average age was 46.00 ($SD = 16.41$, $Mdn = 46$), 66.70% has a higher education degree (1.85% did not finish high school), 31.73% identifies as left-wing (32.85% as centrist; 35.42% as right-wing), 37.72% identifies most as Democrat (29.61% as Independent; 30.28% as Republican), 39.84% checks the news multiple times a day (34.84% once a day; 14.65% weekly; 8.43% less often than weekly; 2.25% never), 54.08% uses social media multiple times a day (23.43% once a day; 9.55% weekly; 5.36% less often than weekly; 7.57% never), 29.11% uses YouTube multiple times a day (22.89% once a day; 26.32% weekly; 16.09% less often than weekly; 5.59% never), and 6.17% uses YouTube for news consumption multiple times a day (12.89% once a day; 16.54% weekly; 23.98% less often than weekly; 40.42% never). In Experiment 1, the

rating task was administered at two different time points: T1 (immediately after watching the video) and T2 (two weeks after watching the video). Participants were randomly assigned to one of six conditions (see Figure 6.1.4.1 for an overview): the short inoculation condition (with posttest at T1 or at T2), the long inoculation condition (with posttest at T1 or T2), or the control condition (with posttest at T1 or T2). The rationale for this design, and specifically for the splitting in different sample groups per posttest time point, is to eliminate repeated testing effects, which could lead to unwanted effect-boosting confounds in the measurement of decay (see e.g., Maertens et al., 2021). Note that "T1" represents the day of the intervention, and as there was no pretest, we refer to"T1" as the posttest at "T1" (unless otherwise specified). This experiment was preregistered on the AsPredicted platform at https://aspredicted.org/WL8_LSK, and all analysis scripts in R, items, and Qualtrics survey files can be found on the OSF repository at https://osf.io/zrq87/?view_only=375c0632fca0444fa07c2bc46a59187b.

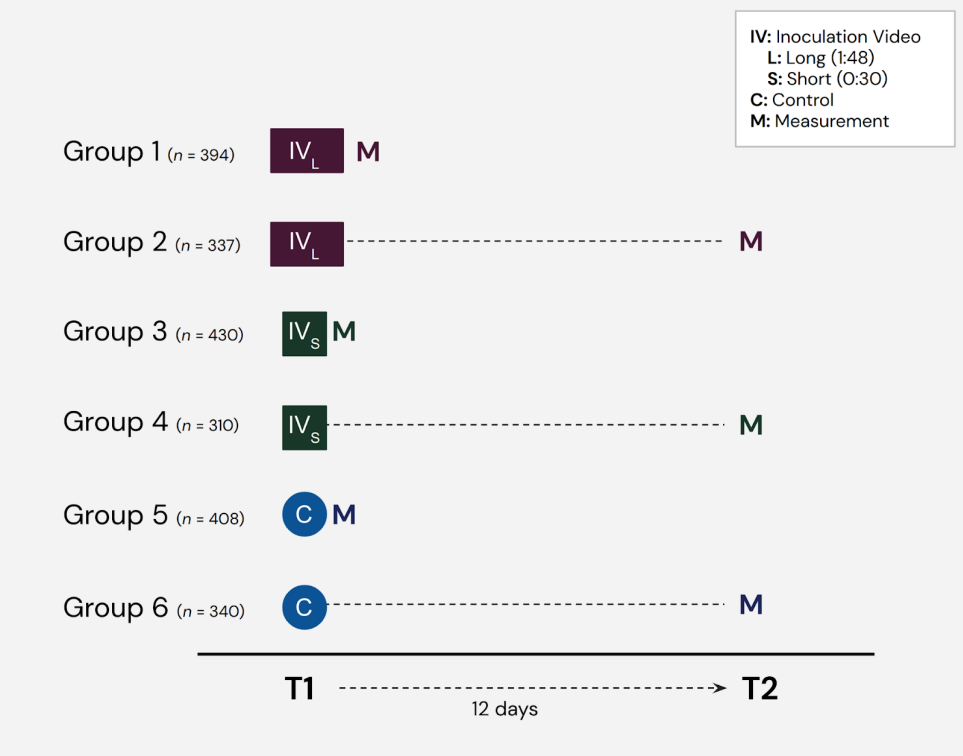

Figure 6.1.4.1. The experimental design of Experiment 1.

Results

Hypothesis Tests

Main Effect of Inoculation. To test our main hypotheses for the "manipulativeness" measure (i.e., manipulative language discernment), we preregistered a two-way (3x2) ANOVA analysis. We found that the omnibus test is significant, $F(5, 2213) = 15.64$, $p < .001$, indicating that we could continue to test our contrasts as planned. As preregistered, we then conducted a series of Tukey-corrected ANOVA contrast tests to test hypotheses H1.1–H1.3. We found that the inoculation effect for the long inoculation video as compared to the control video was significant, $M_{diff} = 0.75$, $t(2213) = 7.43$, $p_{tukey} < .001$, $d = 0.525$, 95% CI [0.385, 0.664], providing evidence in line with H1.1a. Also the short inoculation video compared to the control video led to a significant effect $M_{diff} = 0.63$, $t(2213) = 6.36$, $p_{tukey} < .001$, $d = 0.439$, 95% CI [0.303, 0.575], in line with H1.1b. The above analyses indicated significant medium effect sizes both for the long inoculation video and for the short inoculation video, replicating the original study (Roozenbeek, van der Linden, et al., 2022), in favor of H1.1: both videos significantly improved participants' ability to discern manipulative from non-manipulative content. After establishing the baseline effect, we compared the short and the long inoculation videos and explored the decay over time.

Short vs. Long Videos. We tested the contrast of the manipulative discernment scores after the short and long video. The videos did not show a significantly different effect from one another in terms of T1 effect sizes $M_{diff} = 0.12$, $t(2213) = 1.22$, $p_{tukey} = .826$, $d = 0.085$, 95% CI [-0.051, 0.222], advising rejection of H1.2, indicating that the long and short videos were equally effective in the immediate post-test.

2-Week Effectiveness. Comparing the T2 (*Mdn* = 12 days after T1) and T1 decay in the long inoculation condition, we found that a significant decay takes place, $M_{diff} = -0.36$, $t(2213) = -3.43$, $p_{tukey} = .008$, $d = -0.255$, 95% CI [-0.400, -0.109]. Moreover, after this decay,

the inoculation effect was no longer significantly different from the control condition M_{diff} = 0.24, $t(2213) = 2.23$, $p_{tukey} = .227$, $d = 0.171$, 95% CI [0.020, 0.322]. A similar result could be found when comparing T2 to T1 of the short inoculation videos $M_{diff} = -0.31$, $t(2213) = -2.90$, $p_{tukey} = .044$, $d = -0.216$, 95% CI [-0.362, -0.070], and when comparing T2 short inoculation to T2 control $M_{diff} = 0.18$, $t(2213) = 1.58$, $p_{tukey} = .611$, $d = 0.124$, 95% CI [-0.030, 0.278]. These decay analyses indicated that there is full decay of the inoculation effect when measuring 12 days after T1, leading to the rejection of H1.3.[56] See Figure 6.1.4.2 (Panel A) for a visualization of manipulativeness discernment over time.

Exploratory Analyses

Memory. Although not the focus of this first experiment and not preregistered, we explored the relation between manipulativeness discernment and the objective memory score, and found a significant positive correlation, $t(2217) = 12.15$, $p < .001$, $r = 0.250$, 95% CI [0.210, 0.288]. See Figure 6.1.4.2 (Panel B) for a visualization of inoculation memory over time. As 8 out of 12 questions were yes-or-no questions and therefore had a 50% chance of being correctly answered, the regression to 50% over time observed in the control group can be explained by guessing.

[56] Although not preregistered, we also ran the above analyses with the confidence, trustworthiness, and sharing intent measures. Here, similar to the analyses for manipulativeness, we found significant effects for T1 (each in the expected direction), and significant decay to the extent that the effect is no longer significant when the Tukey *p*-value correction is administered, except for trustworthiness discernment in the long video. A larger sample would be needed to determine the presence of a reduced effect. All effects were driven by the scores for the manipulative items, with minimal change for non-manipulative items.

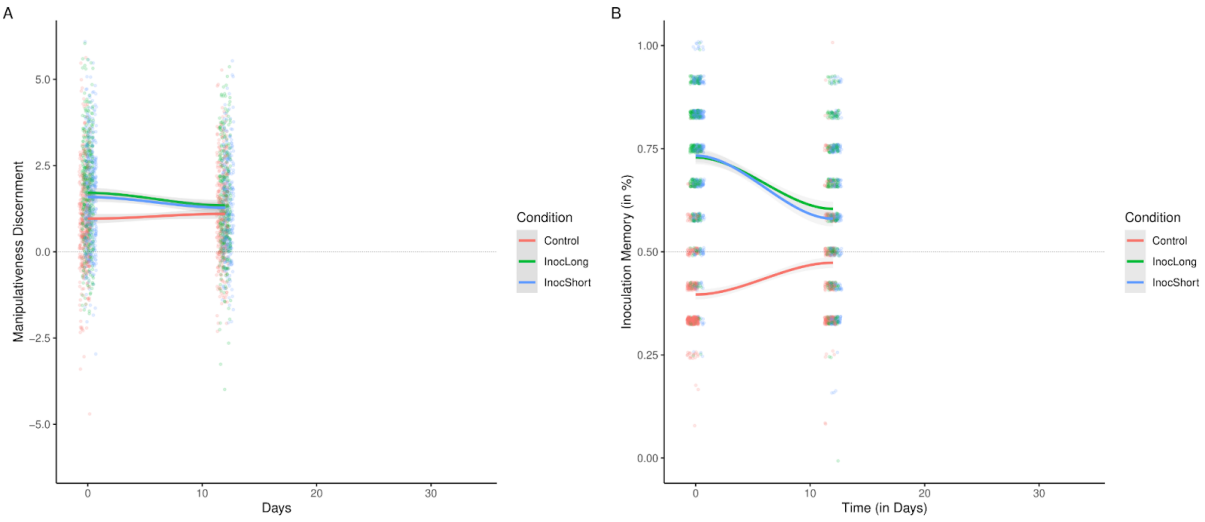

Figure 6.1.4.2. Visual plot of "manipulativeness discernment" (Panel A) and inoculation memory (Panel B) in Experiment 1. Days represents the time elapsed after the intervention. Inoc is the inoculation intervention (Long: 1 minute 48 seconds; Short: 30 seconds). Error bands represent the standard error. $N = 2{,}219$.

Discussion

In a large, preregistered randomized controlled experiment, we investigated the longitudinal effects of video-based inoculation interventions. We found, first of all, that both the longer (1 minute 48 seconds) and shorter (30 seconds) technique-based inoculation videos are highly effective at improving people's ability to spot manipulative social media content, replicating the initial results reported by Roozenbeek, van der Linden, et al. (2022). Second, we found that both videos are equally effective, which is important in light of the scalability of inoculation interventions: for example, 30-second videos can be implemented as non-skippable ads on video sharing platforms, thus significantly increasing their potential reach and impact. Third, we found that 12 days after watching the video, the inoculation effect (which is strong immediately after viewing) has dissipated almost entirely. This finding offers important initial insight into the longevity of video-based inoculations: while we do not yet know what the "decay curve" of the effect looks like, we do know that significant decay

takes before 12 days. Further research is needed to examine the findings from this first decay experiment in more detail. Experiment 2 will bring insight into the shape of the inoculation "decay curve", or when exactly inoculation interventions lose their efficacy after viewing, and explore the mechanisms behind its longevity. Experiment 3, then, will extend Experiment 2 and look at the role of "booster shots" and when to re-administer (or "top up") inoculation interventions after the initial exposure in order to retain their efficacy.

6.1.5 Experiment 2

Methods

Sample and Design

The basis of the video-based inoculation paradigm, including the dependent variables, are the same as in Experiment 1. New in Experiment 2 is that we include only the short videos (0 min 30 sec), have a larger sample size, and include multiple time points (4, 10, and 30 days). In total, we recruited N = 5,191 participants to T1, with random allocation to each condition (see Figure 6.1.5.1 for an overview). After—in line with the preregistration protocol—removing participants that failed both the manipulation and attention checks, participated in the survey more than once, entered the same response to all items of the dependent variable, or did not complete the entire survey, a total of N = 4,821 participants remained. Of our final sample, 51.73% identified as female (47.65% as male; 0.44% as non-binary; 0.10% as "other"; 0.08% preferred not to answer), the average age was 45.79 (SD = 16.46, Mdn = 45), 65.63% has a higher education degree (1.35% did not finish high school), 30.47% identifies as left-wing (35.51% as centrist; 34.02% as right-wing), 35.86% identifies most as Democrat (32.03% as Independent; 29.25% as Republican), 36.57% checks the news multiple times a day (35.72% once a day; 14.87% weekly; 9.83% less often than weekly; 3.01% never), 51.05% uses social media multiple times a day (25.16% once a day; 10.81% weekly; 6.16% less often than weekly; 6.82% never), 27.11% uses YouTube multiple

times a day (22.59% once a day; 27.84% weekly; 16.74% less often than weekly; 5.73% never), and 5.83% uses YouTube for news consumption multiple times a day (11.28% once a day; 15.25% weekly; 22.13% less often than weekly; 45.51% never). Participant attrition levels were lower than the predicted percentages: 24.6% for T2, 28.2% T3, and 39.7% for T4.

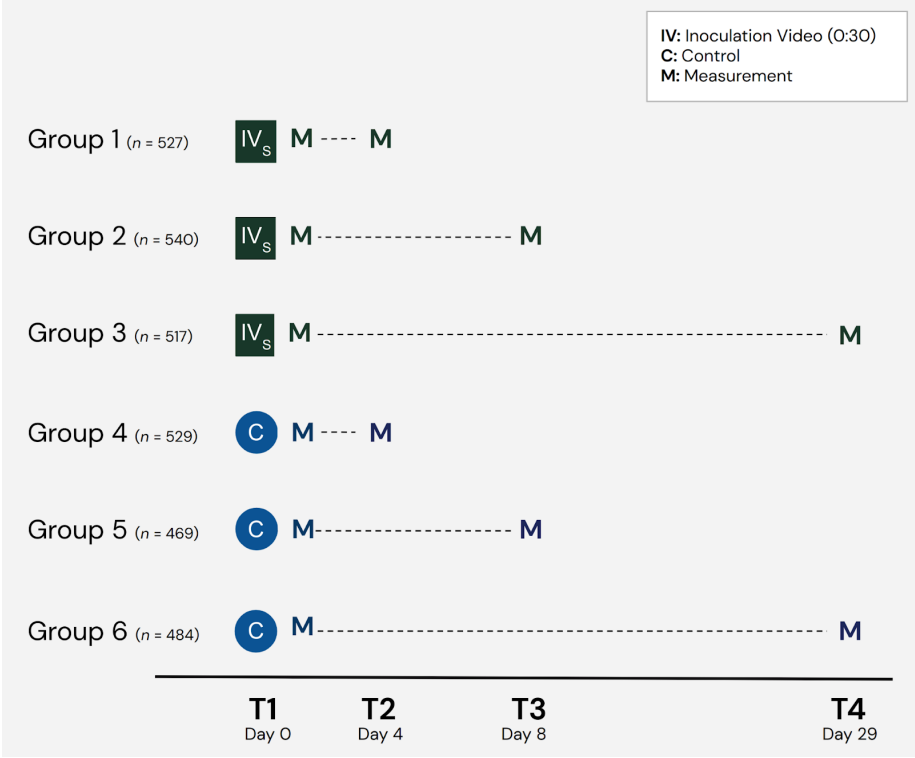

Figure 6.1.5.1. The experimental design of Experiment 2.

Results

Hypothesis Tests

Main Effect. As preregistered, we tested hypothesis H2.1 by running an ANOVA with manipulativeness discernment as the dependent variable and group (inoculated or not) as the independent variable, with the full T1 dataset ($N = 4{,}821$). We found that the ANOVA omnibus test is significant, $F(1, 4819) = 134.73$, $p < .001$. To test H2.1, we looked at the main effect of the intervention at T1 and found that the inoculation effect is significant, $M_{diff} = 0.47$, $t(4819) = 11.61$, $p_{tukey} < .001$, $d = 0.335$, 95% CI [0.278, 0.391].

Decay Analysis. To test the decay hypotheses H2.2, H2.3, and H2.4, we made use of an ANCOVA with T1 discernment as a covariate, post-posttest discernment as a dependent variable, and group and evaluation date as independent variables. In addition, we now used all time points, and only include data from participants who completed the follow-up within 3 days from the intended follow-up date (N = 3,066, $Mdn_{\text{BetweenDays,T2}}$ = 4, $Mdn_{\text{BetweenDays,T3}}$ = 8, $Mdn_{\text{BetweenDays,T4}}$ = 29). The omnibus test was significant, $F(3060) = 12.66$, $p < .001$. In line with our expectations, we found evidence for the stability of the effect over 4 days, with a significant effect compared to the control group, $M_{\text{diff}} = 0.53$, $t(3060) = 6.10$, $p_{\text{tukey}} < .001$, $d = 0.375$, 95% CI [0.254, 0.496], and no significant change in the inoculation groups between the two time points, $M_{\text{diff}} = 0.18$, $t(6124) = 2.54$, $p_{\text{tukey}} = .178$, $d = 0.128$, 95% CI [0.029, 0.226]. After 8 days we found that the effect was still significant compared to the control group, $M_{\text{diff}} = 0.41$, $t(3060) = 4.56$, $p_{\text{tukey}} < .001$, $d = 0.288$, 95% CI [0.164, 0.412], and—contrary to our expectations—that there was no significant change between T1 and T3 in the inoculation groups, $M_{\text{diff}} = 0.04$, $t(6124) = 0.54$, $p_{\text{tukey}} > .999$, $d = 0.027$, 95% CI [-0.070, 0.125]. After 29 days we found that, in line with our preregistered hypothesis, that the inoculation effect is no longer significant compared to the control group, $M_{\text{diff}} = 0.18$, $t(3060) = 2.05$, $p_{\text{tukey}} = .315$, $d = 0.130$, 95% CI [0.006, 0.254], but without a significant decay in the inoculation group when comparing T4 to T1, $M_{\text{diff}} = -0.01$, $t(6124) = 2.05$, $p_{\text{tukey}} > .999$, $d = -0.010$, 95% CI [-0.109, 0.089]. See Figure 6.1.5.2 for a visualization of manipulativeness discernment (Panel A) and memory retention (Panel B) over time for each condition.

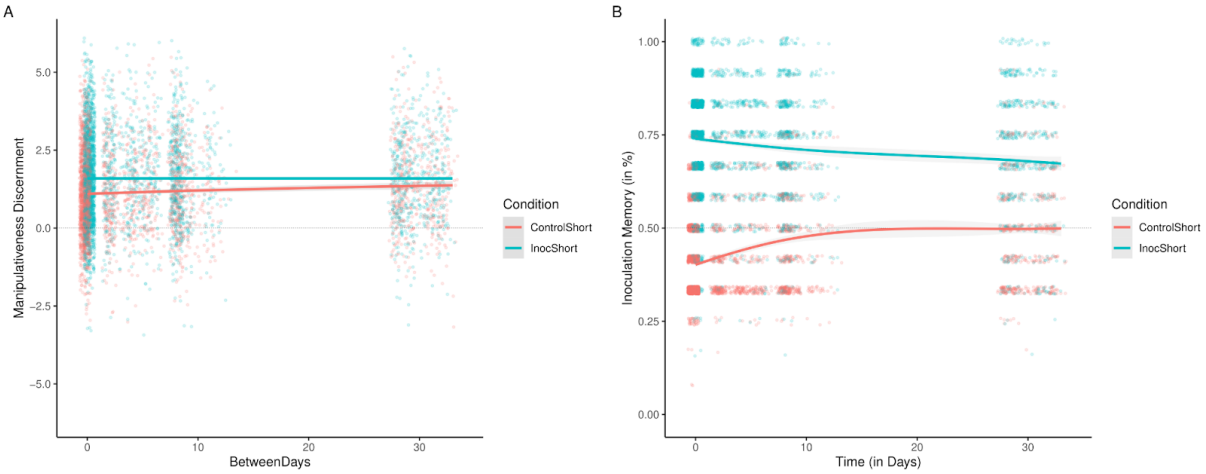

Figure 6.1.5.2. Visual plot of "manipulativeness discernment" (Panel A) and inoculation memory (Panel B) in Experiment 2. Control and InocShort represent the 30-second control and inoculation videos. Days represent the time passed since the inoculation intervention. Confidence bands represent the standard error. $N = 3{,}066$.

Memory and Motivation. To test H2.5 and H2.6 and compare the mechanisms with the results from Study 2 and Study 4, we modeled an SEM model using the *lavaan* package in R (Rosseel, 2012) with second posttest memory and motivational threat as mediators for the manipulativeness discernment at second posttest, and T1 inoculation as the predictor variable, allowing direct effects from inoculation to memory, motivational threat, and discernment, and direct effects from memory and motivational threat to discernment. See Figure 6.1.5.3 for a schematic visualisation of the model and its direct and indirect relationships, and Table 6.1.5.4 for a table with all estimates.

As predicted, we found that memory directly predicts the inoculation effect at a later time point, $t(3062) = 7.78$, $p < .001$, $\beta = 0.169$, 95% CI [0.126, 0.212], as did motivation, $t(3062) = 7.85$, $p < .001$, $\beta = 0.138$, 95% CI [0.104, 0.173]. As can be seen in Table 6.1.5.4, all indirect and all component effects were significant with a significant total effect of the inoculation intervention, $t(3062) = 7.32$, $p < .001$, $\beta = 0.262$, 95% CI [0.192, 0.333], and no

significant direct effect of the intervention, $t(3062) = 1.07$, $p = .238$, $\beta = 0.047$, 95% CI [-0.038, 0.131], providing evidence for full mediation.

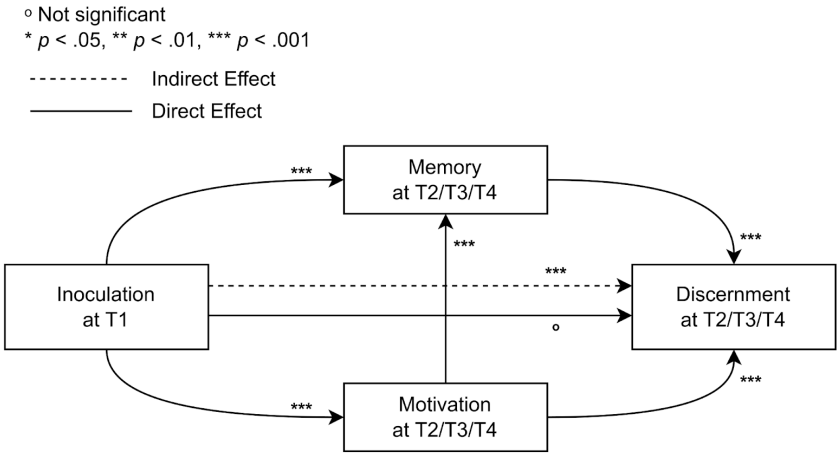

Figure 6.1.5.3. The memory-motivation model of inoculation in Experiment 2 ($N = 3{,}066$).

Table 6.1.5.4

*Memory-Motivation Model Estimates in Study 5, Experiment 2 (*N *= 3,066)*

Effect	*z*	*p*	β	95% CI		*SE*
				LL	*UL*	
Indirect						
Inoc.T1 ⇒ Memory.T234 ⇒ Discernment.T234	**7.638**	**< .001**	**0.197**	**0.147**	**0.248**	**0.026**
Inoc.T1 ⇒ Motivation.T234 ⇒ Discernment.T234	**3.136**	**.002**	**0.017**	**0.006**	**0.028**	**0.005**
Inoc.T1 ⇒ Motivation.T234 ⇒ Memory.T234 ⇒ Discernment.T234	**2.576**	**.010**	**0.001**	**0.000**	**0.002**	**0.001**
Component						
Inoc.T1 ⇒ Memory.T234	**39.974**	**< .001**	**1.167**	**1.110**	**1.224**	**0.029**
Memory.T234 ⇒ Discernment.T234	**7.782**	**< .001**	**0.169**	**0.127**	**0.212**	**0.022**
Inoc.T1 ⇒ Motivation.T234	**3.420**	**< .001**	**0.123**	**0.053**	**0.194**	**0.036**
Motivation.T234 ⇒ Discernment.T234	**7.853**	**< .001**	**0.138**	**0.104**	**0.173**	**0.018**
Motivation.T234 ⇒ Memory.T234	**4.531**	**< .001**	**0.066**	**0.038**	**0.095**	**0.015**
Direct						
Inoc.T1 ⇒ Discernment.T234	1.073	.283	0.047	-0.038	0.131	0.043
Total						
Inoc.T1 ⇒ Discernment.T234	**7.322**	**< .001**	**0.262**	**0.192**	**0.333**	**0.036**

Discussion

In Experiment 2 we investigated the decay of the video-based inoculation effect, including its decay curve and the mechanisms behind it. We found that the inoculation effect remained significant till 8 days after the intervention. Moreover, when comparing the results to Experiment 1, we see an indication that an immediate posttest could have served as an inoculation "booster shot" itself. In other words, a "booster shot" in the form of an immediate posttest may help to establish the effect for a longer period of time, even when not repeated at a later time point, which is compatible but different from the repeated testing booster shots found by Maertens et al. (2021). After 29 days, the intervention effect was no longer significant. Investigating the mechanisms of decay revealed that both inoculation memory and motivational threat are predictors of manipulativeness discernment after 29 days, with memory showing a descriptively stronger effect size. This experiment also corroborated evidence from Experiment 1 that the short inoculation video is an effective intervention in the short term, and that it may provide beneficial to explore "booster videos" in a third experiment.

6.1.6 Experiment 3

Methods

Sample and Design

In Experiment 3 we built further on the design of Experiment 2, as well as Study 2 and Study 4, by combining multiple videos to test booster effects over time. In this final study we aimed to test and disentangle the two effects that drive inoculation effects: the threat component, and the refutational preemption (Compton, 2013, 2021). All participants were exposed to two different videos, a first video at T1, and a second video at T2 (*Mdn* = 9 days later). The first video was either the control video or the short inoculation used in Experiment 2. The second video was the same control or inoculation video repeated, a "threat booster"

video focused on increasing levels of threat and motivation (Booster A), or a "memory booster" video focused on reminding people of what they learned in the original intervention (Booster B). We designed Booster A (the threat booster video) in such a way that it employed emotional music and warned people about manipulative online content, but it did not explain the methods that are used to mislead people nor use any of the content from the original video (i.e., only threat, no refutational preemption). Booster B on the other hand omitted the emotional music and affective forewarnings, but it did repeat the explanation of the techniques that can be used to mislead people using emotional language with similar content to the original video. Finally, all participants took the manipulativeness discernment test at T1 and at T3 (*Mdn* = 29 days later). This allowed us to disentangle and link effects at immediate posttest and at later posttest to enable testing the memory-motivation model. All participants were randomly allocated to the different video combinations (see Figure 6.1.6.1 for an overview).

In total, we recorded 6,164 survey responses at T1. As preregistered, we excluded incomplete and low-quality responses, leading to a T1 sample size of 5,703. Finally, we removed participants that did not participate in all three parts of the survey or did not participate in the follow-up sessions within 3 days before or after the intended time (T2: 10 days after, T3: 30 days after). This led to a final sample size of 2,220, with an average of 444 participants per group. This is slightly below but close to the intended 548 participants per group (participant attrition from T1 to T3 was 61%, slightly above the estimated 55%). In our final sample, 55.14% identified as female (44.50% as male; 0.23% as non-binary; 0.09% as "other"; 0.05% preferred not to answer), the average age was 53.29 (*SD* = 14.48, *Mdn* = 55), 67.48% has a higher education degree (1.40% did not complete high school), 29.19% identifies as left-wing (34.23% as centrist; 36.58% as right-wing), 36.13% identifies most as Democrat (28.87% as Independent; 32.07% as Republican), 40.90% checks the news

multiple times a day (36.85% once a day; 12.52% weekly; 7.70% less often than weekly; 2.03% never), 45.68% uses social media multiple times a day (25.09% once a day; 10.90% weekly; 7.21% less often than weekly; 11.13% never), 21.89% uses YouTube multiple times a day (19.77% once a day; 29.59% weekly; 21.08% less often than weekly; 7.66% never), and 5.72% uses YouTube for news consumption multiple times a day (9.86% once a day; 11.49% weekly; 21.62% less often than weekly; 51.31% never).

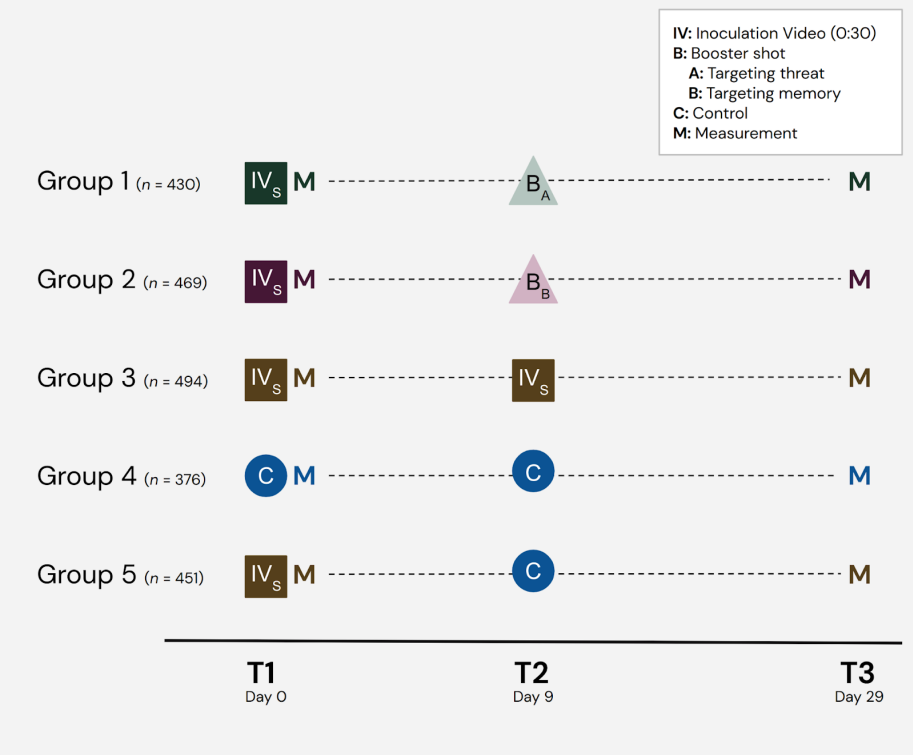

Figure 6.1.6.1. The experimental design of Experiment 3.

Results

Hypothesis Tests

Main Effect. We tested H3.1 by exploring the main effect of the short inoculation video at T1, with the full sample size ($N = 5{,}703$), according to the preregistered protocol. A one-way ANOVA analysis with inoculation status (control group vs. inoculated group) as the predictor and manipulativeness discernment as the outcome variable. The omnibus test was

significant, $F(1, 5701) = 76.81$, $p < .001$, indicating we can look at our specific main effect analysis. We found that the effect is significant with a similar strength as observed in the previous experiment, $M_{diff} = 0.44$, $t(5701) = 8.76$, $p_{tukey} < .001$, $d = 0.295$, 95% CI [0.229, 0.361].

Decay Analysis. We tested the hypotheses with relation to the inoculation effect longevity (H3.2, H3.3, H3.4, H3.5) by using a repeated measures ANOVA analysis with group and test (T1, T3) as predictor variables and manipulativeness discernment as the dependent variable. After confirming a significant omnibus test, $F(4, 2215) = 10.14$, $p < .001$, we looked, as preregistered, at the contrasts between the groups at T3. We found, contrary to our hypothesis H3.2, that the group which has not seen a repeated inoculation video or any of the two booster videos still showed a significant inoculation effect at T3 (29 days after T1), $M_{diff} = 0.34$, $t(2215) = 3.30$, $p_{tukey} = .009$, $d = 0.230$, 95% CI [0.093, 0.367]. In line with our predictions for H3.3, H3.4, and H3.5, we found that the inoculation effects remained significant for the groups that were boosted at T2 (9 days after T1), whether it was through a repetition of the inoculation, $M_{diff} = 0.36$, $t(2215) = 3.65$, $p_{tukey} = .003$, $d = 0.250$, 95% CI [0.115, 0.384], a "threat booster" video, $M_{diff} = 0.38$, $t(2215) = 3.66$, $p_{tukey} = .002$, $d = 0.258$, 95% CI [0.120, 0.397], or a "memory booster" video, $M_{diff} = 0.64$, $t(2215) = 6.35$, $p_{tukey} < .001$, $d = 0.440$, 95% CI [0.303, 0.576]. Descriptively, the memory booster video performs the best, with 100% retention of the original effect size, while the other two booster conditions retain ~86% of the original effect size, and the control booster condition 78%. See Figure 6.1.6.2 for a visual plot of the manipulativeness discernment (Panel A) and the memory retention (Panel B) in each condition and their inoculation effect over time in Experiment 3.

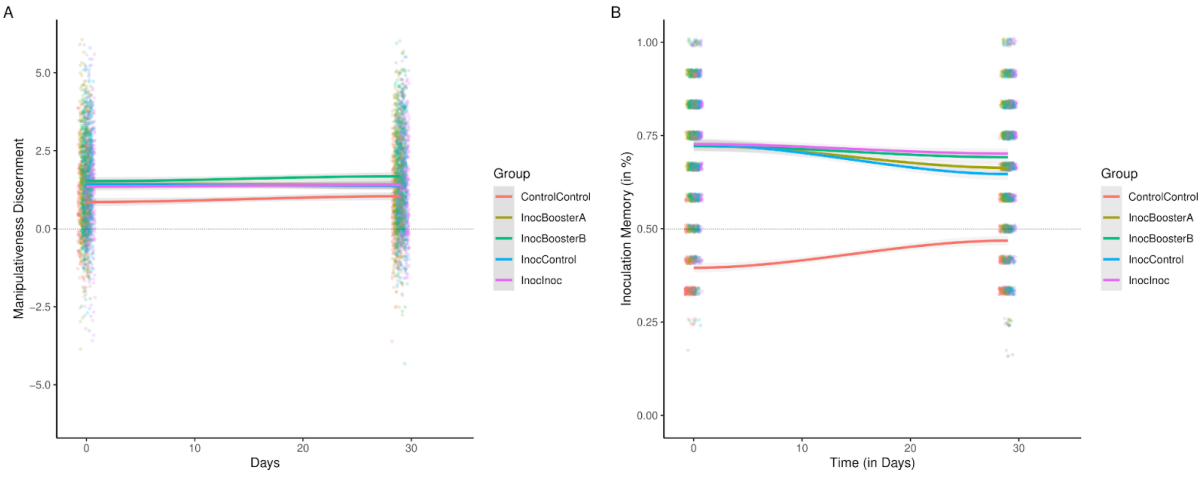

Figure 6.1.6.2. Panel A is a visual plot of "manipulativeness discernment" in Experiment 3 (N = 2,220). Panel B presents the inoculation memory retention for each group. Days represent the time passed since the inoculation intervention. Confidence bands represent the standard error.

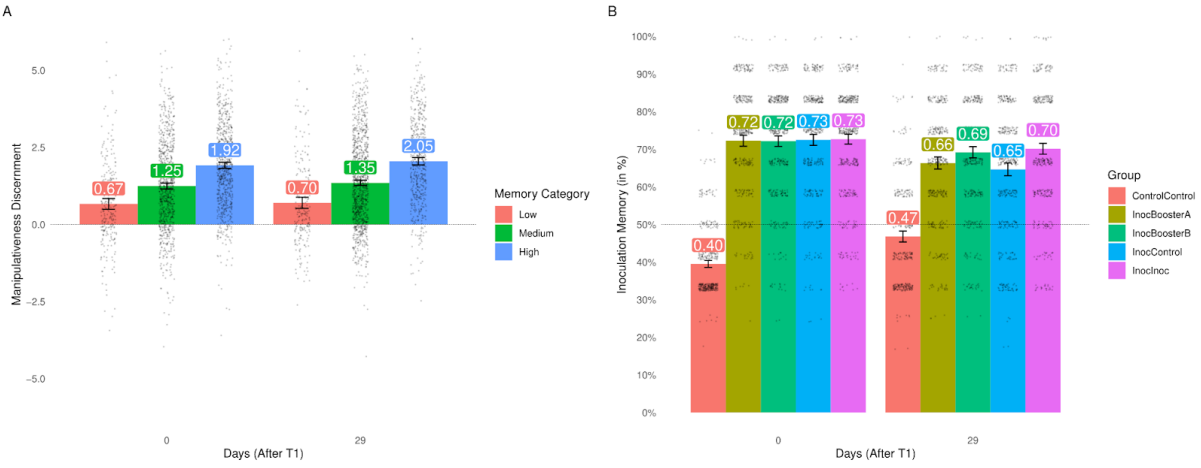

Figure 6.1.6.3. Panel A represents the discernment scores for participants in the inoculation condition, split by memory and test date (N = 1,844). Panel B represents memory of the inoculation intervention, split by condition and test date (N = 2,220). Error bars represent 95% Confidence Intervals.

Memory and Motivation. After our decay hypotheses, we investigated the effect of booster sessions on the memory and motivation variables (H3.6, H3.7, H3.8, H3.9). The first three hypotheses were tested using the same repeated measures ANOVA analysis but now with motivation (a) or memory (b) as the outcome variables. Model a, $F(4, 2215) = 8.41$, $p = .003$, and model b, $F(4, 2215) = 132.04$, $p < .001$, both showed a significant omnibus test. Looking at the preregistered contrasts, we found that a threat-focused booster video (Booster A) did not have a significant impact on motivation, $M_{diff} = 0.03$, $t(2215) = 0.28$, $p_{tukey} = .999$, $d = 0.019$, 95% CI [-0.113, 0.151], nor on memory, $M_{diff} = 0.20$, $t(2215) = 1.51$, $p_{tukey} = .556$, $d = 0.102$, 95% CI [-0.030, 0.234]. Neither the re-inoculation procedure, $M_{diff} = 0.09$, $t(2215) = 0.97$, $p_{tukey} = .870$, $d = 0.063$, 95% CI [-0.065, 0.191], nor the memory-focused booster video (Booster B), $M_{diff} = 0.17$, $t(2215) = 1.81$, $p_{tukey} = .366$, $d = 0.120$, 95% CI [-0.010, 0.249], had a significant effect on motivation. Meanwhile both the re-inoculation procedure, $M_{diff} = 0.66$, $t(2215) = 5.09$, $p_{tukey} < .001$, $d = 0.331$, 95% CI [0.203, 0.460], and the memory-focused booster video (Booster B), $M_{diff} = 0.54$, $t(2215) = 4.14$, $p_{tukey} < .001$, $d = 0.273$, 95% CI [0.143, 0.403], had a significant effect on memory.

Finally, to test H9, we implemented a SEM model in the *lavaan* R package (Rosseel, 2012) similar to Study 2 and Study 4, to test whether the effects of the intervention on the outcome variable are mediated by motivation and memory. We found evidence for partial mediation at T1, with memory, $b = 0.22$, $t(2216) = 13.24$, $p < .001$, $\beta = 0.351$, 95% CI [0.299, 0.403], and motivation, $b = 0.08$, $t(2216) = 3.89$, $p < .001$, $\beta = 0.079$, 95% CI [0.039, 0.118], having an effect on manipulativeness discernment, but meanwhile keeping intact a direct effect of inoculation, $b = 0.32$, $t(2216) = 3.13$, $p = .002$, $\beta = 0.220$, 95% CI [0.082, 0.358]. At T3 we found full mediation, with inoculation no longer being significant directly, $b = 0.05$, $t(2216) = 0.526$, $p = .599$, $\beta = 0.031$, 95% CI [-0.085, 0.148], but memory, $b = 0.17$, $t(2216) = 11.58$, $p < .001$, $\beta = 0.249$, 95% CI [0.215, 0.303], and motivation, $b = 0.16$, $t(2216) = 7.77$, p

< .001, β = 0.158, 95% CI [0.118, 0.198], having a remaining influence, whilst inoculation directly influenced memory, $b = 2.50$, $t(2217) = 22.11$, $p < .001$, β = 1.133, 95% CI [1.032, 1.233], and motivation, $b = 0.28$, $t(2217) = 3.44$, $p < .001$, β = 0.194, 95% CI [0.084, 0.305]. See Figure 6.1.6.3 for an overview of the inoculation effect for each memory category (Panel A) and a bar graph of the memory scores (Panel B) for each time point in Experiment 3.

The data in this study allowed us to go one step further in our SEM analyses than Study 2 and Study 4 allowed, as due to the immediate posttest *and* a second posttest at a later date, we now have longitudinal data for the mapping of paths between time points. To test the memory-motivation model in its entirety, we therefore created an SEM model that includes inoculation at T1, memory at T1 *and* T3, motivation at T1 *and* T3, and the booster interventions at T2. See Figure 6.1.6.4 for a simplified visual representation of the memory-motivation SEM model at T3 and Table 6.1.6.5 for the complete model estimates. As can be seen from the estimates provided in the table and the visual summary, the video inoculation effects work indirectly via memory and motivation, with the largest effects for memory, both for inoculation on memory, and for memory on manipulativeness discernment performance. The role of motivation seems to be particularly important for the T1 memory formation and relatedly, the motivation booster (Booster A) presented at T2 did not provide any additional benefits for performance or motivation at T3. Meanwhile, the memory booster (Booster B) presented at T2 successfully managed to boost the inoculation effect at T3 by boosting the inoculation memory, which in turn was the best predictor for the effect retention at T3. These findings are in line with the memory-motivation model of inoculation.

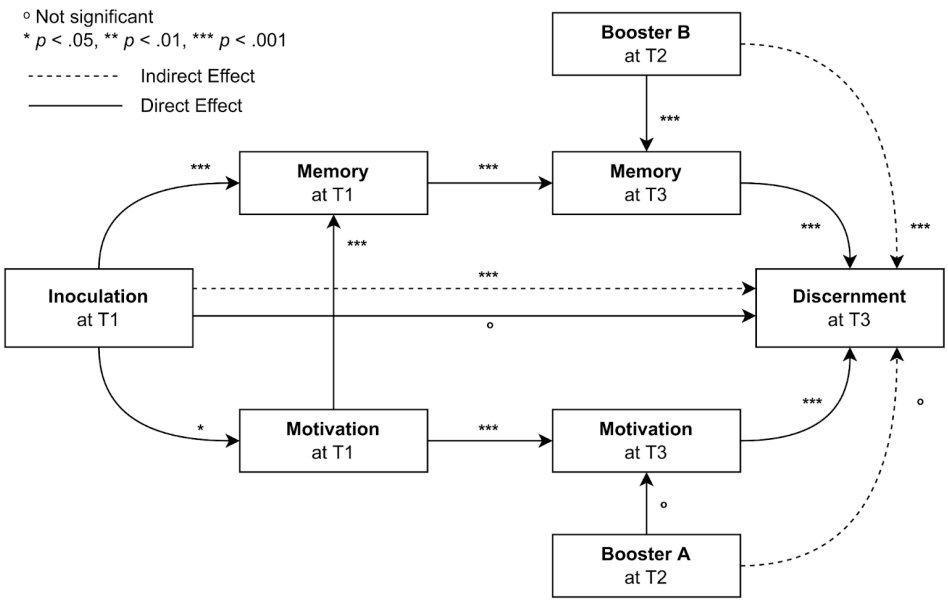

Figure 6.1.6.4. The memory-motivation model of inoculation in Experiment 3 (N = 2,220).

Table 6.1.6.5

*Memory-Motivation Model Estimates in Study 5, Experiment 3 (*N *= 2,220)*

Effect	*z*	*p*	β	95% CI		*SE*
				LL	*UL*	
Indirect						
BoosterA ⇒ Motivation.T3 ⇒ Discernment.T3	1.149	.251	0.009	-0.006	0.024	0.008
Inoc2 ⇒ Motivation.T3 ⇒ Discernment.T3	0.835	.403	0.006	-0.008	0.020	0.007
Inoc2 ⇒ Memory.T3 ⇒ Discernment.T3	**5.567**	**< .001**	**0.074**	**0.048**	**0.099**	**0.013**
BoosterB ⇒ Memory.T3 ⇒ Discernment.T3	**4.883**	**< .001**	**0.064**	**0.038**	**0.089**	**0.013**
Inoc1 ⇒ Motivation.T1 ⇒ Motivation.T3 ⇒ Discernment.T3	**2.345**	**.019**	**0.011**	**0.002**	**0.020**	**0.005**
Inoc1 ⇒ Memory.T1 ⇒ Memory.T3 ⇒ Discernment.T3	**10.497**	**< .001**	**0.228**	**0.186**	**0.271**	**0.022**
Inoc1 ⇒ Motivation.T1 ⇒ Memory.T1 ⇒ Memory.T3 ⇒ Discernment.T3	**2.129**	**.033**	**0.001**	**0.000**	**0.003**	**0.001**
Component						
BoosterA ⇒ Motivation.T3	1.162	.245	0.056	-0.039	0.152	0.049
Motivation.T3 ⇒ Discernment.T3	**7.764**	**< .001**	**0.156**	**0.117**	**0.196**	**0.020**
Inoc2 ⇒ Motivation.T3	0.840	.401	0.039	-0.052	0.129	0.046
Inoc2 ⇒ Memory.T3	**6.316**	**< .001**	**0.284**	**0.196**	**0.372**	**0.045**
Memory.T3 ⇒ Discernment.T3	**11.787**	**< .001**	**0.259**	**0.216**	**0.302**	**0.022**
BoosterB ⇒ Memory.T3	**5.365**	**< .001**	**0.245**	**0.156**	**0.335**	**0.046**
Inoc1 ⇒ Motivation.T1	**2.471**	**.013**	**0.140**	**0.029**	**0.250**	**0.056**
Motivation.T1 ⇒ Motivation.T3	**26.356**	**< .001**	**0.488**	**0.452**	**0.525**	**0.019**
Inoc1 ⇒ Memory.T1	**40.009**	**< .001**	**1.718**	**1.634**	**1.802**	**0.043**
Memory.T1 ⇒ Memory.T3	**28.238**	**< .001**	**0.513**	**0.477**	**0.549**	**0.018**
Motivation.T1 ⇒ Memory.T1	**4.546**	**< .001**	**0.073**	**0.042**	**0.105**	**0.016**
Direct						
BoosterA ⇒ Discernment.T3	0.017	.986	0.001	-0.124	0.126	0.064
Inoc2 ⇒ Discernment.T3	-1.095	.274	-0.068	-0.189	0.054	0.062
BoosterB ⇒ Discernment.T3	**2.002**	**.045**	**0.126**	**0.003**	**0.248**	**0.063**
Inoc1 ⇒ Discernment.T3	-0.652	.515	-0.045	-0.180	0.090	0.069
Total						
BoosterA ⇒ Discernment.T3	0.416	.677	0.028	-0.103	0.159	0.067
Inoc2 ⇒ Discernment.T3	0.297	.766	0.019	-0.107	0.146	0.065
BoosterB ⇒ Discernment.T3	**3.179**	**.001**	**0.208**	**0.080**	**0.336**	**0.065**
Inoc1 ⇒ Discernment.T3	**3.302**	**< .001**	**0.228**	**0.093**	**0.364**	**0.069**

Exploratory Analyses

Mechanisms of Inoculation. We performed a dominance analysis to investigate the most dominant predictors of the inoculation effect at T3, and found that memory (41% dominance) and motivational threat (27%) were the best predictors of inoculation longevity. To further demonstrate the role of memory in inoculation, we looked at the effect of the

inoculation intervention for people who have a good memory of the intervention at T3, and found a large effect, $M_{diff} = 1.01$, $t(896) = 10.76$, $p_{tukey} < .001$, $d = 0.728$, 95% CI [0.591, 0.865], for manipulativeness at 29 days, while only a small effect was found for those with an average memory of the intervention, $M_{diff} = 0.31$, $t(1471) = 3.68$, $p_{tukey} < .001$, $d = 0.220$, 95% CI [0.102, 0.337]. See Table 6.1.6.6 for an overview of the dominance analysis outcome and Figure 6.1.6.7 for a plot depicting the role of memory (Panel A) next to the role of motivation (Panel B). For a plot combining the data from the three experiments demonstrating the role of memory at each time point, see Figure 6.1.6.8.

Table 6.1.6.6

Dominance Analysis in Experiment 3, at T3, N = *2,220*

Variable	Dominance
Memory	41%
Motivational Threat	27%
Fear	10%
Apprehensive Threat	9%
Issue Involvement	8%
Issue Talk	2%
Issue Accessibility	2%
Self-Reported Remembrance	0%

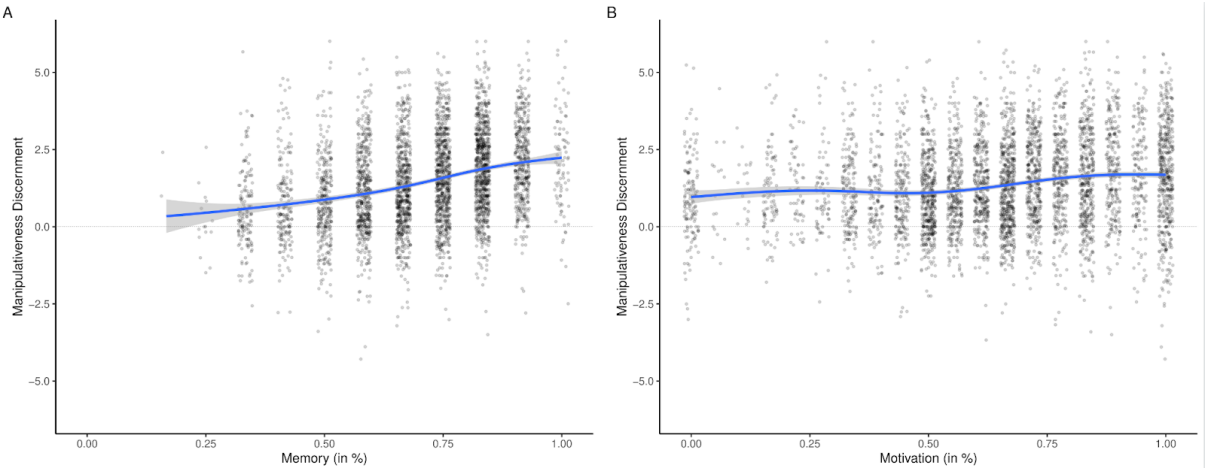

Figure 6.1.6.7. Plot of memory (Panel A) and motivation (Panel B) in relation to the inoculation effect ($N = 2{,}220$).

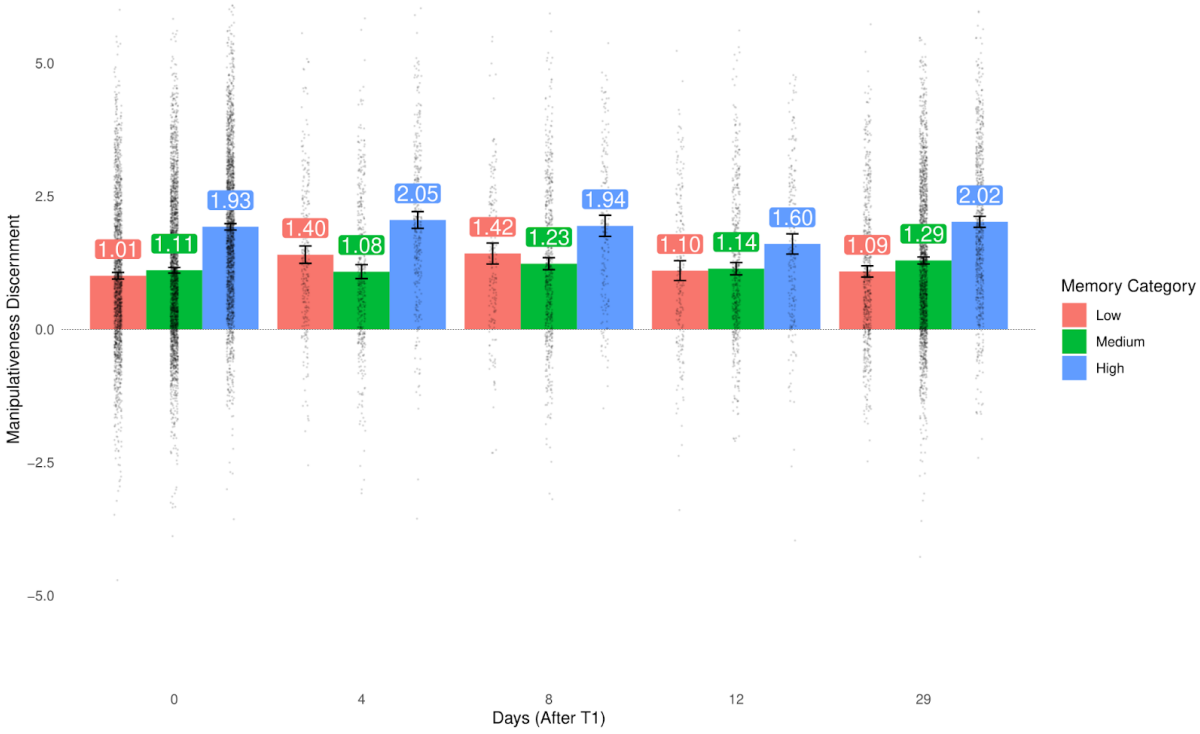

Figure 6.1.6.8. The inoculation effect separated by inoculation memory recall and time after the intervention (in days) in the combined sample ($N_{datapoints}$ = 12,791).

Political Leaning. As a robustness check, we investigated whether the inoculation effect is significant across varying political leanings, and found that the inoculation effect shows a similar pattern for each subgroup. See Figure 6.1.6.9 for a visualization of the inoculation effect across political ideology subgroups in the combined sample of the three experiments.

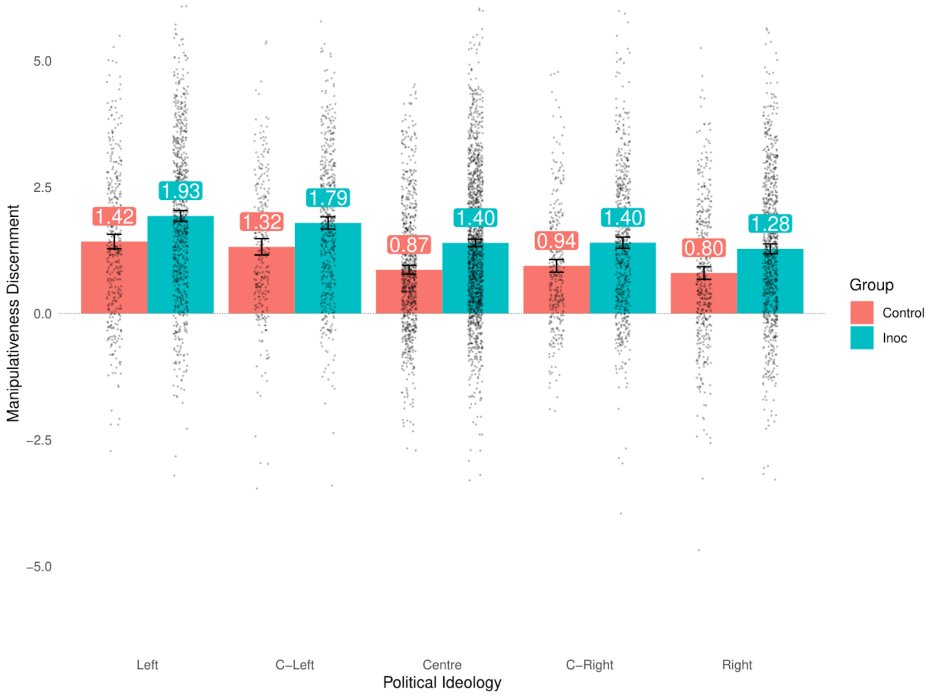

Figure 6.1.6.9. Inoculation effect across political leaning in the combined sample ($N_{datapoints}$ = 6,518). Error bars represent the 95% confidence interval.

Concept Mapping. As a final exploratory analysis we used a concept mapping question as a qualitative analysis of participants' memory recall. Participants were asked to write down concepts they remembered from the original video in an open box, before they received directed memory questions. We used natural language processing packages in R (*tm*, Theußl et al., 2012 ; *SnowballC*, Bouchet-Valat, 2020; *wordcloud*, Fellows et al., 2018) to clean the text data and mapped the data by counting the frequency of the words entered. The results of this question show that the inoculated groups have a distinct memory network at T3, showing that even before prompting participants with direct memory questions, they were able to recall key concepts of the inoculation intervention (e.g., "emotional", "manipulative", "language"). The control group participants recalled concepts from the control video (e.g., "eye", "macular", "degeneration"). See Figure 6.1.6.10 for a word cloud comparison of the T3 (29 days) responses in the control condition compared to the memory booster condition.

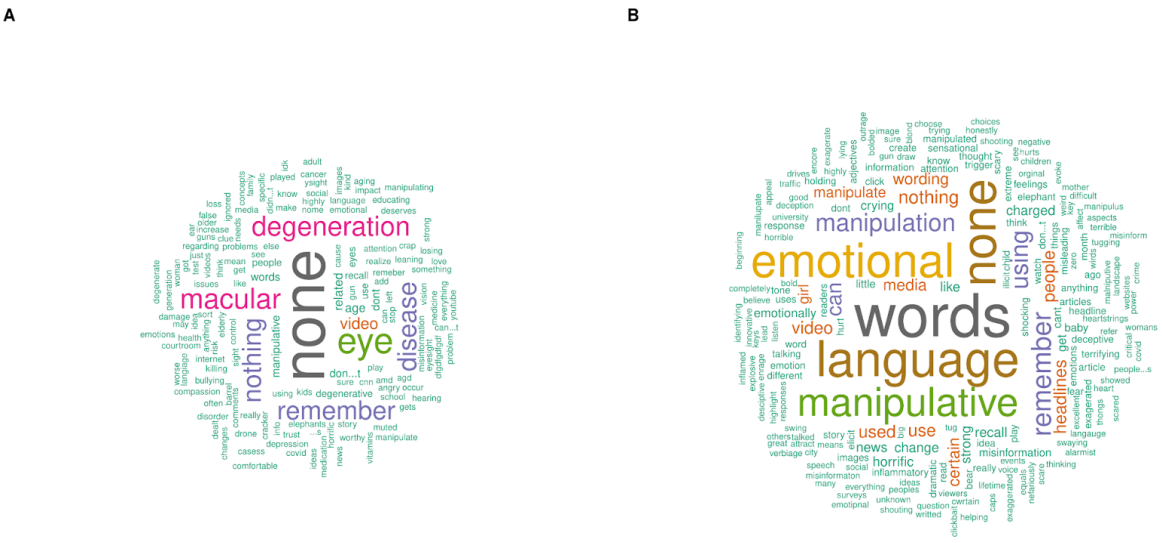

Figure 6.1.6.10. Word cloud of memory recall question responses at T3 for participants in the control group (Panel A) and in the inoculation group that received a memory booster (Panel B). Larger words represent a higher occurrence of the word. $N = 626$.

Discussion

In Experiment 3 we set out to expand on our findings on the longevity of inoculation effects and test repeated inoculation videos and two types of booster videos to extend its longevity further. We found that the inoculation effect remained significant after a period of 29 days even without any booster intervention, which was contrary to our expectations. Descriptive analyses show that a repetition of the original inoculation intervention—in line with Ivanov et al. (2018)—and a memory-based booster presented 9 days after the intervention, could be effective at strengthening the effect of the intervention, while threat-based booster videos may not be as effective as videos that refreshed the memory about the "emotional language" techniques explained in the original inoculation video. Investigating the underlying mechanisms, we found that both memory and motivation were robust predictors of the inoculation effect at both T1 and T3, and served as mediators for the T1 intervention effect at T3 (after a period of 29 days). Of all the measured explanatory

mechanisms, memory was the strongest and most dominant predictor for the inoculation effect over time, followed by motivational threat in second place.

6.1.7 Discussion

In a large, preregistered, randomized controlled study, we investigated the longitudinal effects of video-based inoculation interventions. We found, first of all, that both the longer (1:48 min) and shorter (0:30 min) technique-based inoculation videos are effective at improving people's ability to discern manipulative social media content from neutral social media content, in a direct replication of the initial results reported by Roozenbeek, van der Linden, et al. (2022). The effectiveness of the short inoculation video was robust across all experiments, with Experiment 3, which had the largest T1 sample size (N = 5,703), showing an effect of d = 0.295, 95% CI [0.229, 0.361] for the short video. Across all three experiments (d_{Exp1} = 0.439, d_{Exp2} = 0.335, d_{Exp3} = 0.295) we find a small-to-medium effect, similar to the meta-analytic inoculation effect size d = 0.43 (Banas & Rains, 2010). Furthermore, these effects were robust across the different political leanings of the participants. In addition, the finding that both videos are equally effective is important in light of the scalability of inoculation interventions: for example, 30-second videos can be implemented as non-skippable ads on video sharing platforms, thus significantly increasing their potential reach and impact. While earlier work has explored inoculation videos (Lim & Ki, 2007; Nabi, 2003; Pfau et al., 1992, 2000; Varker & Devilly, 2012; Vraga et al., 2021), this work shows the robustness of a short, video-based intervention, of which its longer (and similarly effective) variant has been shown to over 5 million YouTube users with positive results (Roozenbeek, van der Linden, et al., 2022). In addition, in contrast to the previous four studies, we utilized manipulativeness discernment as the main outcome variable as well as varying manipulativeness-to-neutral item ratios. This means that in this study we have an indicator not just of manipulativeness detection, but also of how people judge misleading

social media items in contrast to neutral items. This shows that inoculation effects can reflect not only scepticism, but also a skill to distinguish content more accurately.

Second, we found that in the first experiment, two weeks after watching the video, the inoculation effect (which is strong immediately after viewing) had dissipated to a level of insignificance, with a remaining non-significant inoculation effect size of $d = 0.124$, 95% CI [-0.030, 0.278]. This finding is similar to the result of Zerback et al. (2021), who also found that the inoculation effect can decay almost entirely after two weeks. We could not replicate these findings in Experiment 2, $d = 0.130$, 95% CI [0.006, 0.254], and Experiment 3, $d = 0.230$, 95% CI [0.093, 0.367], where unexpectedly, we found significant effects up to 29 days after the intervention—although the remaining effect sizes were small. This finding is more in line with Ivanov et al. (2018), who also found significant effects 4 weeks after an initial inoculation intervention. While future research will have to explore the discrepancy between the first experiment and the other two experiments, we hypothesise that this is due to a change in experimental design. In Experiment 1, participants did not complete an immediate posttest after the intervention, and thus only ever received one posttest discernment task. Meanwhile, in Experiments 2 and 3, participants received an immediate posttest, as well as a posttest at a later time point. It is known from previous research that the actual testing of participants can serve as a booster (Maertens et al., 2021), and thus future research should try to disentangle whether this also means that an immediate posttest could have such influence.

In all three experiments, we found an important role of *memory*. In Experiment 3, we found an indirect effect of inoculation through memory of $\beta = 0.228$, 95% CI [0.186, 0.271], while the indirect effect of inoculation through motivation was only $\beta = 0.011$, 95% CI [0.002, 0.020]. We also found that the inoculation effect becomes weaker over time, but remains intact for those who remember the inoculation video. These findings demonstrate the importance of potential booster treatments, as well as of designing engaging and memorable

materials that people enjoy interacting with to make it more likely that they remember the intervention.

Finally, we offer new insight into the theoretical question on the role of threat versus memory in inoculation effects and their longevity decay (Compton, 2013): do psychological inoculations stop being effective because the sense of threat (of an impending attack on one's beliefs) disappears (Compton, 2021), or because people forget what they learned (Maertens et al., 2021; Pfau et al., 2005)? Our findings provide support for both, but with a more important role for memory, finding that people with the same memory of the content of the inoculation intervention showed almost equal discernment performance at each time point and across the experiments. Nevertheless, although memory of the inoculation was a better predictor, motivational threat remained an important variable in predicting variance in inoculation effects immediately after the intervention and at later time points, in line with findings by Banas and Richards (2017) and Richards and Banas (2018). In Experiment 3 we went one step further and investigated the role of booster interventions, including a threat-focused booster video (Booster A) and a technique reiteration booster (Booster B). The results indicate that motivation may be important for memory formation at T1, and that a memory booster at T2 can help to boost inoculation effects over time via improved memory. The T2 memory booster (Booster B) had a small effect on memory at T3, $\beta = 0.245$, 95% [0.156, 0.335], while the T2 threat booster (Booster A) did not manage to increase motivational threat, $\beta = 0.056$, 95% CI [-0.039, 0.152], and also had no significant effect on the outcome measure. Future research will have to explore whether this is because threat-based booster videos are not effective, or whether it was specifically the threat-based booster video that we designed that was not effective at doing so. While this provides new insights into the potential role of boosters in inoculation, we agree with other scholars that still *"more needs to be learned about the best way to structure and time booster messages"* (Ivanov & Parrott,

2017, p. 23). Thus, although our main video-based intervention did manage to increase motivational threat, the threat-focused booster video did not. These findings however all fit within the proposed memory-motivation model of the inoculation effect (see Chapter 2), and will inspire future experiments to explore the interaction between motivation and memory formation (learning).

Chapter 7

Methodological Issues

Working on the different inoculation paradigms revealed the theoretical and methodological challenges that come with the measurement of inoculation effects and misinformation susceptibility. In this chapter, I explore the encountered methodological issues and discuss potential solutions. In the first study (**Study 6**) I explore whether the evaluation of inoculation interventions might be influenced by whether or not a pre-test is administered (testing effects), as well as what the impact is of the choice of specific item sets in inoculation designs (item effects). The study shows that a pre-test has only a limited to no influence on the measurement of inoculation effects, but that the choice of items *does* have an impact. All materials, supplementary materials, analysis scripts, and clean and raw datasets for this study can be found on the OSF repository at https://doi.org/10.17605/OSF.IO/FGEQJ.

Exploring the literature for solutions made clear that there is a need for a standardised theoretical and empirical framework on misinformation susceptibility. The issue reaches beyond my own research—for all its momentum, the current proliferation of misinformation research across disciplines has one big caveat that is as much theoretical as it is psychometric: it is not clear what researchers mean by susceptibility to misinformation and it is even less clear how to cleanly measure it. Similarly, while research has often focused on selecting the right misinformation stimuli, limited attention has been given to the selection of high-quality real news items. In a final study (**Study 7**), I set out to develop a psychometrically validated measurement instrument that puts equal emphasis on real news and fake news, the Misinformation Susceptibility Test (MIST), as well as a new theoretical measurement framework based on disambiguating the various aspects of resilience to misinformation. All supplementary materials, analysis scripts in R (incl. cleaning code), clean

and raw datasets, preregistrations, item lists, and Qualtrics survey files can be found on the OSF repository for this study at https://osf.io/r7phc/.

Study 6

7.1 Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation: Solomon Revisited[57]

7.1.1 Abstract

Online misinformation is a pervasive global problem. In response, psychologists have recently explored the theory of psychological inoculation: if people are preemptively exposed to a weakened version of a misinformation technique, they can build up cognitive resistance. This study addresses two unanswered methodological questions about a widely adopted online "fake news" inoculation game, *Bad News*. First, research in this area has often looked at pre- and post-intervention difference scores for the same items, which may imply that any observed effects are specific to the survey items themselves. Second, it is possible that using a pre-test influences the outcome variable of interest, or that the pre-test may interact with the intervention. We investigate both item and testing effects in two online studies using the *Bad News* game. For the item effect, we examine if inoculation effects are still observed when different items are used in the pre- and post-test. To examine the testing effect, we use a Solomon's Three Group Design. We find that inoculation interventions are minimally influenced by item effects, and not by testing effects. We show that inoculation interventions are effective at improving people's ability to spot misinformation techniques and that the *Bad News* game does not make people more sceptical of real news. We discuss the larger relevance of these findings for evaluating real-world psychological interventions.

[57] Study 6 has been published as *"Disentangling Item and Testing Effects in Inoculation Research on Online Misinformation: Solomon Revisited"* in Educational and Psychological Measurement (Roozenbeek, Maertens, et al., 2021). It was written in collaboration with Dr Jon Roozenbeek (University of Cambridge), Dr William McClanahan (Max Planck Institute for the Study of Crime, Security, and Law), and Professor Sander van der Linden (University of Cambridge). The first three authors (Dr Jon Roozenbeek, Rakoen Maertens, and Dr William McClanahan) contributed equally to this study.

7.1.2 Introduction

The spread of online misinformation is a threat to democracy and a pervasive global problem that is proving to be tenacious and difficult to eradicate (Lewandowsky, Ecker, & Cook, 2017; van der Linden & Roozenbeek, 2020; World Economic Forum, 2018). Part of the reason for this tenacity can be found in the complexity of the problem: misinformation is not merely information that is provably false, as this classification would unjustly target harmless content such as satirical articles. Misinformation also includes information that is manipulative or otherwise harmful, e.g., through misrepresentation, leaving out important elements of a story, or deliberately fuelling intergroup conflict by exploiting societal or political wedge issues, without necessarily having to be blatantly "fake" (Tandoc et al., 2018; Roozenbeek & van der Linden, 2019a, 2019b). Efforts to combat misinformation have included introducing or changing legislation (Human Rights Watch, 2018), implementing detection algorithms (Ozbay & Alatas, 2020), promoting fact-checking and "debunking" (Nyhan & Reifler, 2012), and developing educational programmes such as media or digital literacy (Carlsson, 2019). Each of these solutions have advantages as well as important disadvantages, such as issues surrounding freedom of speech and expression (Ermert, 2018; Human Rights Watch, 2018), the disproportionate consequences of wrongly labelling or deleting content (Hao, 2018; Pennycook et al., 2020; Pieters, 2018), the limited reach and effectiveness of media literacy interventions (Guess et al., 2019; Livingstone, 2018), the "continued influence effect" of misinformation once it has taken hold in memory (Lewandowsky et al., 2012), and the fact that misinformation may spread further, faster, and deeper on social media than other types of news, thus ensuring that fact-checking efforts are likely to remain behind the curve (Vosoughi et al., 2018).

Accordingly, researchers have increasingly attempted to leverage basic insights from social and educational psychology to find new and *preemptive* solutions to the problem of

online misinformation (Fazio, 2020; Roozenbeek, Nygren, & van der Linden, 2020). One promising avenue in this regard is inoculation theory (Compton, 2012; McGuire & Papageorgis, 1961a; McGuire, 1964; van der Linden, Leiserowitz, et al., 2017; van der Linden, Maibach, et al., 2017), often referred to as the "grandfather of resistance to persuasion" (Eagly & Chaiken, 1993, p. 561). Inoculation theory posits that it is possible to build cognitive resistance against future persuasion attempts by pre-emptively introducing a weakened version of a particular argument, much like a "real" vaccine confers resistance against a pathogen (McGuire & Papageorgis, 1961b). Although meta-analyses have supported the efficacy of inoculation interventions (Banas & Rains, 2010), only recently has research begun testing inoculation theory in the context of misinformation.

A notable example of a real-world inoculation intervention against online misinformation is the award-winning *Bad News* game[58], an online browser game in which players take on the role of a fake news creator and actively generate their own content. The game simulates a social media feed and players see short texts or images, and can react to them in a variety of ways (see Figure 7.1.2.1 for a screenshot of the user interface). Their goal is to acquire as many followers as possible while also building credibility for their fake news platform. Through humour (Compton, 2018) and perspective-taking, players are warned and exposed to severely weakened doses of common misinformation techniques in a controlled learning environment in an attempt to help confer broad-spectrum immunity against future misinformation attacks.[59]

[58] The game is free and accessible online via www.getbadnews.com.
[59] For a detailed description of the game and the various misinformation techniques participants learn about, please see Roozenbeek and van der Linden (2019).

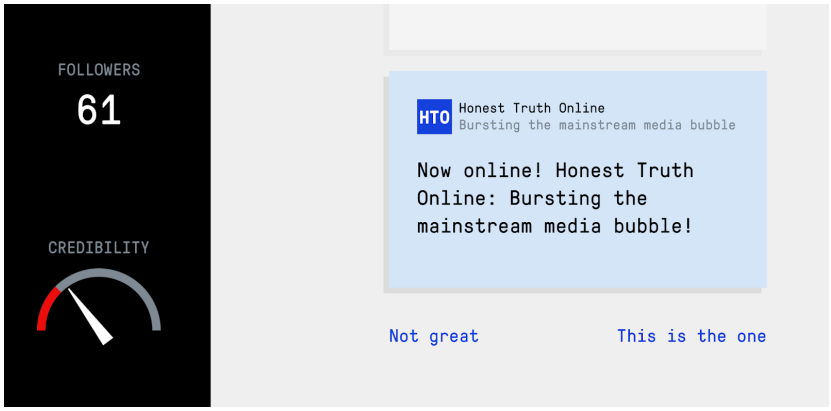

Figure 7.1.2.1. A screenshot of the in-game user interface of *Bad News*.

Although several studies have shown that the *Bad News* game can successfully improve players' ability to spot misinformation techniques (e.g., Roozenbeek & van der Linden, 2019a; Basol et al., 2020), including cross-cultural evaluations (Roozenbeek et al., 2020)[60], several key open questions about how to *measure* the effectiveness of game-based inoculation interventions remain unanswered. Specifically, the effectiveness of *Bad News* has so far been assessed by looking at pre-post differences (within-subject) and difference-in-differences scores between groups (i.e., the *Bad News* game versus a control task) in which the items (in the form of 'fake' and 'real' Twitter posts) used for the pre-test and post-test are the same.[61] However, presenting the same items twice (in the pre- and post-test) may indicate a learning effect specific to the items themselves, rather than of participants' latent ability to spot misinformation in headlines that they have never seen before (i.e., an 'item effect') .

Moreover, by simply assessing a construct (i.e., a pre-test), researchers may inadvertently influence the outcome variable of interest, or the initial assessment may interact with the intervention and moderate its influence (Song & Ward, 2015). Statistician Andrew Gelman refers to this potential issue as "poisoning the well" (Gelman, 2017). For example by

[60] The *Bad News* game is currently playable online in 15 different languages including German (www.getbadnews.de), Dutch (www.slechtnieuws.nl), Russian (www.getbadnewsrussian.com), Ukrainian (www.getbadnewsukraine.com), Swedish (www.badnewsgame.se) and Esperanto (www.misinformado.net).
[61] Importantly, the test items are different from the "training set" the players are exposed to during gameplay.

taking a practice test (i.e., a pre-test) students may memorise some answers to questions, directly influencing the outcome variable at a second assessment (i.e., a post-test). Additionally, once pre-tested, the same students may also become more or less comfortable with the testing process, which in turn may moderate the influence of an intervention aimed at improving their scores (i.e., a 'testing effect').

Accordingly, this study addresses both item and testing effects in the context of a real-world intervention. Specifically, we examine if the use of a pre-test influences the effects of inoculation against misinformation in two ways. First, to investigate the item effect, we examine if inoculation effects are still observed when different items are used in the pre- and post-test. To examine the testing effect, we use a Solomon Three-Group Design (Solomon, 1949), in which participants are randomly assigned to one of three groups. Group 1 can be considered a traditional experimental group with a pre-test, intervention (the *Bad News* game), and a post-test. Group 2 participates in a pre-test and post-test, without an active intervention (the control group). Finally, Group 3 receives the intervention and the post-test, but no pre-test. This allows us to isolate the unique influence of the pre-test, intervention, and interaction of the pre-test and intervention (pre-test X intervention) on the mean difference score between pre and post-test.

7.1.3 Methods

Experimental Design

We present the results of two separate pre-registered[62] experiments. Following the approach laid out by Roozenbeek and van der Linden (2019), we implemented voluntary in-game surveys both at the start and at the end of the *Bad News* game. After being introduced to the game mechanics, players of the game were asked to participate in a

[62] Pre-registration was done via AsPredicted.org for both the item effect study (https://aspredicted.org/kt4gm.pdf) and the testing effect study (https://aspredicted.org/wy59x.pdf). Some minor alterations were made as compared to the pre-registration, which are described in Supplementary Declarations S1 and S2.

scientific study. If players provided informed consent, we recorded their responses to a series of pre-post test items (in the form of 'fake' or 'real' Twitter posts) as well as several demographic questions (see the "outcome measures" section for more details). The studies were approved by the Psychology Research Ethics Committee, and both experiments were run within the game over a period of six months (from 30 June 2019 to 17 December 2019). The full R scripts, materials, and datasets are available via the OSF: https://doi.org/10.17605/OSF.IO/FGEQJ.

As this study uses an intervention (the *Bad News* game) that is: 1) an online experiment with consistent measurements and instruments, 2) is short in duration (the game takes about 15 minutes to complete), 3) relies on relatively large sample sizes (N_1 = 480, N_2 = 1,679), and 4) ensures that all participants are randomly assigned a condition, traditional threats to internal validity such as history effects, maturation effects, instrumentation effects, regression to the mean, participant selection bias[63], and systematic attrition are minimised.

Experiment 1: Item Effect

In Experiment 1, to investigate the item effect, two different sets, each consisting of six fake (one per misinformation technique) and two 'real' (control) items (a total of 8 items per set) were used, which we will call Set A and Set B. Table S1 provides a full overview of both item sets. After a series of demographic questions, we randomly presented participants with either Set A or Set B in the pre-test. In the post-test after gameplay, participants who had seen Set A in the pre-test were shown Set B in the post-test, and vice versa. If there is no item effect, participants should rate tweets containing misinformation as significantly less reliable after gameplay, even if the pre-test and post-test items are different. This leads to the following hypotheses (see also Figure 7.1.3.1 and Table 7.1.3.3 below):

[63] We note that because the game is free to play, and that individuals must "opt-in" to be a part of the academic study, there is potential for a self-selection bias (Campbell, 1957) at two points: at the decision to play the game and then again at the decision to "opt into" the research. In addition, we were not able to control for participants' country of origin, as consistent with GDPR guidelines and our ethics application, the in-game survey could not ask for this information.

To test if the inoculation effect remains intact:

[H1] When comparing an index of the same fake items, pre-test (group x) to post-test (group y), there is a decrease in the perceived reliability of misinformation, but not for real news. **[H1a]** Set A-A. **[H1b]** Set B-B.

[H2] When comparing an index of different fake items, pre-test (group x) to post-test (group x), there is a decrease in perceived reliability of misinformation, but not for real news. **[H2a]** Set A-B. **[H2b]** Set B-A.

To test if the inoculation effect changes:

[H3] There are no significant difference-in-differences of changes in perceived reliability of news items between Set A-B and Set B-A.

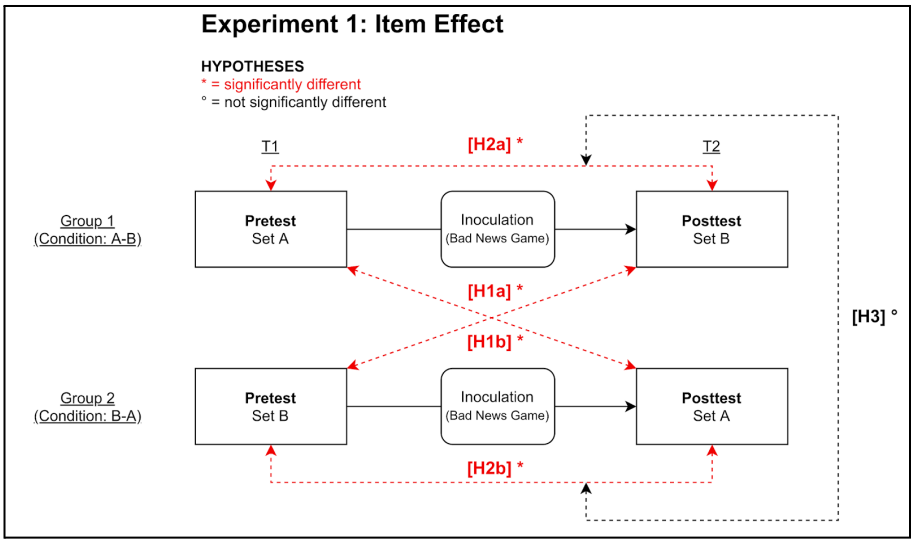

Figure 7.1.3.1. Flowchart with hypotheses for the *Item Effects* experiment.

Experiment 2: Testing Effect

Experiment 2 investigated the testing effect using three different in-game surveys (in line with Solomon's Three Group Design)[64]: 1) a standard experimental group which first answered demographic questions, then did a pre-test, then played the *Bad News* game, and then did a post-test (Group 1)[65]; 2) a traditional control group (without an intervention) in which participants were shown a pre-test, then answered a series of demographic questions, and were then shown a post-test, after which they continued with the remainder of the *Bad News* game (Group 2); and 3) a post-test-only group, which first answered demographic questions (without a pre-test), then played the *Bad News* game, and then did a post-test (Group 3).[66] If there is no testing effect, the inoculation effect should be the same when a pre-test is administered compared to no pre-test (i.e., when comparing Group 1 to Group 3). This leads to the following hypotheses (see Figure 7.1.3.2 and Table 7.1.3.3):

To test the total effect:

[H1] A total effect is observed when comparing mean pre-test perceived reliability ratings to mean post-test reliability ratings in the standard experimental group (Group 1).

[64] We recognize that Solomon also proposed a four-group design, with the fourth group only receiving the post-test. However, in Solomon's design, this fourth group is used to assess environmental factors that could explain mean difference scores between pre- and post-test due to factors such as a time delay or exposure to certain events. As discussed by Solomon (1949), since there is a minimal time delay between pre and post-test in our study (approximately 15 minutes), the influence of the environment may be considered zero. As such, we did not deem it necessary to include a fourth group in this study.

[65] Throughout this paper, Groups 1, 2 and 3 refer to Solomon's (1949) original three groups.

[66] The treatment groups received item Set A with the *polarization* fake news item and the *brands* real news item of Set B, the control group received the complete Set B. This was the result of an error in the implementation of the in-game survey, but should have minimal influence on our results, as only the within-group differences are calculated for this group. To control for confounds because of this discrepancy, we have looked at the same analysis performed only with the shared items between all conditions, and find similar results. We therefore do not see this as a major limitation to the results presented here. See Supplementary Analysis S2 and Supplementary Figure S5 for a detailed overview.

To test the pre-test effect:

[H2] No effect is observed when comparing the mean pre-test perceived reliability rating to the mean post-test reliability rating in the control group (Group 2).

To test if the pre-test interacts with the intervention:

[H3] There is no significant difference between the mean post-test reliability rating of Group 1 subtracted by the pre-test effect and the mean post-test reliability ratings in the post-test-only experimental group (Group 3).

To test the corrected inoculation effect:

[H4] There is a significant difference between the difference in fake news reliability ratings of the pre-test and the post-test in Group 1, when the pre-test effect and the interaction effect are subtracted from the post-test mean, but not for real news.

To confirm the inoculation effect through alternative analysis:

[H5] An inoculation effect is observed when comparing the mean pre-test reliability rating of Group 1 to the mean post-test reliability rating in Group 3.

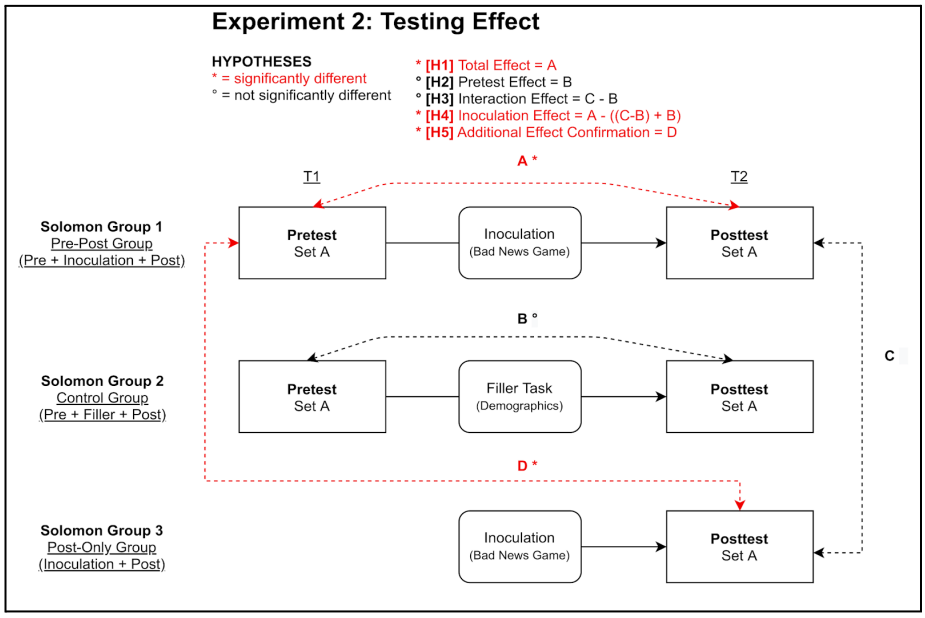

Figure 7.1.3.2. Flowchart with hypotheses for the *Testing Effects* experiment.

Table 7.1.3.3
Experimental setup

Experiment	Condition	Solomon	Demographics first	Pre-test	Intervention	Post-test
Item Effect	A-B		Yes	Set A	Bad News	Set B
	B-A		Yes	Set B	Bad News	Set A
Testing Effect	Pre-Post	Group 1 (E_1)	Yes	Set A	Bad News	Set A
	Control*	Group 2 (C_1)	No	Set A	Demographics	Set A
	Post only	Group 3 (E_2)	Yes	-	Bad News	Set A

* 6 out of 8 items were different in the control group as compared to the other two groups in this experiment; see last footnote.

Participants

Prior power analysis for the detection of effect sizes of d = -0.30 (based on Roozenbeek and van der Linden, 2019) with a Bonferroni corrected α of .01, indicated that for each study we needed 536 participants per group (two groups for Experiment 1, three groups for Experiment 2) for a power of .99. In total, we collected 36,966 responses from participants who started the study in the *Bad News* game[67]. After removing duplicate cases and filtering on complete cases only, 2,182 unique participants remained who completed the full experiment (480 for Experiment 1 and 1,679 for Experiment 2). Due to two back-end

[67] We recognize that the sample is self-selected, as it only contains participants who visited the *Bad News* website.

technical errors, data had to be recollected for Experiment 1. The first two data collection attempts (N_1 = 2,408, N_2 = 1,532) were unsuccessful due to errors in the implementation of the survey; specifically, the first attempt did not include the polarization item, and the second attempt contained the same polarization item in both Set A and Set B. All technical errors were eventually fixed, allowing for successful data collection, albeit smaller in sample size as result of collection limitations. Supplementary Table S4 shows the full results in detail.

The final sample consisted of 51% men, 59% of participants were between the age of 18 and 29, 57% identified as liberal (1-7 Likert scale), and 48% indicated having completed higher education. In addition, 76% of participants used social media regularly or daily; 60% indicated they use Twitter; and 78% checked the news either regularly or every day. A full overview of the sample demographics can be found in Supplementary Table S2.

Outcome Measures

As mentioned above, two sets of Twitter posts were designed (Set A and Set B). Each set of items contained a total of 8 Twitter posts: two 'real' (control) tweets (that do not contain any misinformation technique; e.g., *President Trump wants to build a wall between the US and Mexico*), and six 'fake' tweets that contain misinformation (one for each technique; e.g., *The Bitcoin exchange rate is being manipulated by a small group of rich bankers #InvestigateNow* for the "conspiracy" technique; see Roozenbeek & van der Linden 2019a; for a more detailed description). The 6 misinformation techniques used in these tweets are: impersonation, emotion, polarization, conspiracist ideation, discrediting opponents, and trolling (Basol et al., 2020; Roozenbeek & van der Linden, 2019a; Roozenbeek et al., 2020). The items were designed to be balanced as well as realistic, but not "real" in the sense that they constitute real-life examples of fake (verifiably false) news. Following Roozenbeek and van der Linden (2019), we chose this approach (rather than using real-life examples of fake news) for several reasons: 1) to avoid memory confounds (e.g., people may have seen a fake

news story before); 2) to better be able to isolate each misinformation technique; 3) to balance the items for political neutrality; and 4) to avoid using only "fake" information so as to also include other manipulation techniques (without being explicitly false). To give an example: the two items making use of the "polarization" technique are *Clear difference in career success between left-wing and right-wing voters #Promotion* (Set A) and *The myth of equal IQ between left-wing and right-wing people exposed #TruthMatters* (Set B). Supplementary Table S1 contains the full list of items.

The primary dependent measure in both studies was participants' ability to recognize misleading content in the form of simulated Twitter posts that made use of one (or none in the case of a control item) of the six misinformation techniques learned in *Bad News.* To meet this aim, participants were asked to rate the reliability of each Twitter post on a 7-point Likert scale (see Figure 7.1.3.4 for a typical example)[68].

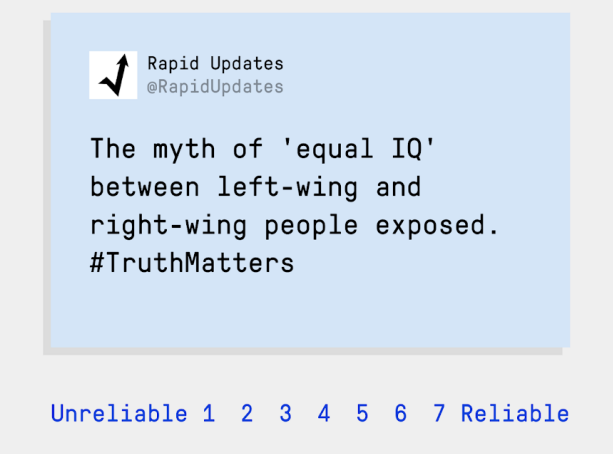

Figure 7.1.3.4. Example item.

[68] Throughout this paper, we refer to this outcome measure as "reliability judgments" following Basol et al. (2020) and Roozenbeek and van der Linden (2019). Here, we do not refer to "reliability" in the technical sense of the internal consistency of psychological measurements (e.g., see Aldridge et al., 2017), but rather in the literal sense, as the perceived reliability of misinformation.

7.1.4 Results

Experiment 1: Item Effects[69]

Fake News

In Experiment 1 we investigated whether the use of different item sets, compared to using the same items for the pre-test and post-test, influences the inoculation effect. Participants had to rate the reliability of each item ($M_{\text{fake,SetA}}$ = 2.73, SD = 1.26, Cronbach's α = 0.74; $M_{\text{fake,SetB}}$ = 2.60, SD = 1.10, Cronbach's α = 0.65) Descriptively, we found a decrease in reliability from pre-test to post-test for both groups ($M_{\text{diff,GroupAB}}$ = -0.22, $SE_{\text{diff,GroupAB}}$ = 0.07; $M_{\text{diff,GroupBA}}$ = -0.35, $SE_{\text{diff,GroupBA}}$ = 0.08), which may also be represented as a shift of the distribution (see density plots in Figure 7.1.4.1, panel B). However, descriptive and visual analyses also suggest an unequal starting point for the different item sets ($M_{\text{pre,SetA}}$ = 2.73, $SD_{\text{pre,SetA}}$ = 1.26; $M_{pre,SetB}$ = 2.60, $SD_{pre,SetB}$ = 1.10), and a differential effectiveness for each item set ($M_{\text{diff,SetA}}$ = -0.49, $SE_{\text{diff,SetA}}$ = 0.11; $M_{\text{diff,SetB}}$ = -0.08, $SE_{\text{diff,SetB}}$ = 0.12)[70]. Figure 7.1.4.1 shows a visualization of the results in bar graphs (Panel A) and density plots (Panel B).

[69] The AsPredicted.org pre-registration can be found here: https://aspredicted.org/kt4gm.pdf, all alterations are explained in Supplementary Declaration S1.
[70] A table with raw means can be found in Supplementary Table S3.

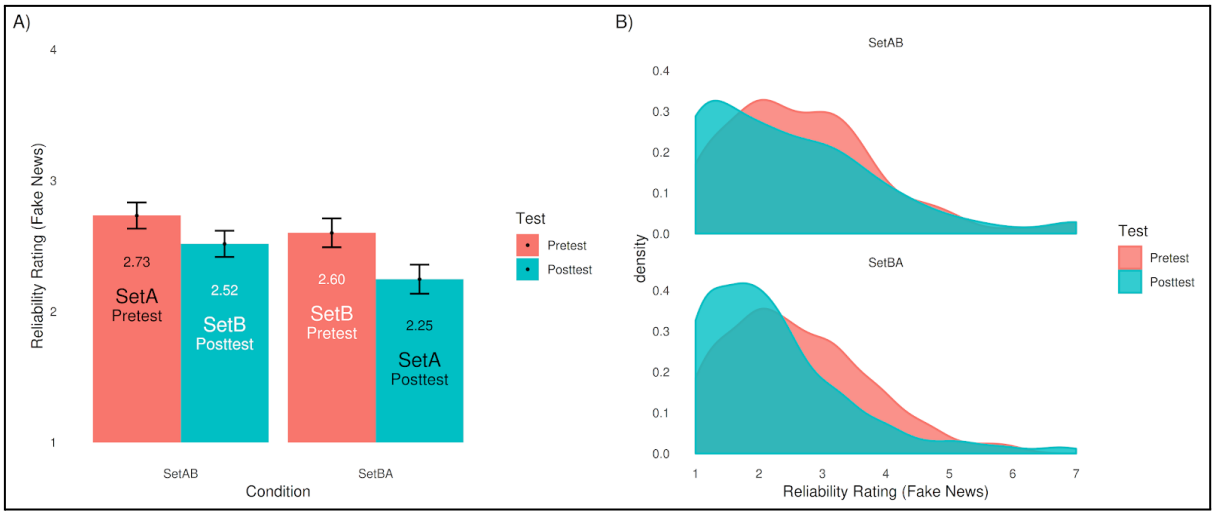

Figure 7.1.4.1[71]. Bar chart (A) and density plots (B) of fake news reliability ratings in the *Item Effects* study. $N = 480$. Error bars represent 95% Confidence Intervals.

To put this to hypothesis testing, we performed five tests, accounting for multiple testing using a Bonferroni correction ($\alpha = .05/5$ tests $= .01$). We first looked at the hypothesis tests investigating whether we still find an inoculation effect using the new procedure of crossing item sets. We tested if the inoculation effect persists when crossing the items between groups, in order to compare Set A (B) pre-test measures to Set A (B) post-test measures. We found a significant effect for Set A (pre-test Group 1) vs Set A (post-test Group 2) **[H1a]** ($M_{diff} = -0.49$, 95% CI_M [-0.70, -0.27], $t(474) = -4.39$, $p < .001$, $d = -0.401$, 95% CI_d [-0.583, -0.218]), but no significant difference for Set B (pre-test Group 2) vs Set B (post-test Group 1) **[H1b]** ($M_{diff} = -0.08$, 95% CI_M [-0.31, 0.14], $t(447) = -0.72$, $p = .47$, $d = -0.066$, 95% CI_d [-0.245, 0.113]). To investigate whether this equals the absence of any effect of interest (Lakens et al., 2018, 2020), we conducted an equivalence test using two one-sided tests (TOSTs).[72] We confirmed statistical equivalence to zero for Set B – Set B ($t(447) = 2.57$, $p = .005$), with the Smallest Effect Size of Interest (SESOI) as $d = (-)0.30$ and $\alpha = 0.01$, in line

[71] For a reversed plotting, where grouping is organized per Set rather than per group, see Supplementary Figure S3.

[72] With a lower SESOI ($d = 0.23$; based on the effect size found for the Set B-Set B test; see Supplementary Table S4), statistical equivalence is no longer equal to zero, suggesting that the preregistered expected effect size of $d = 0.30$ was too high. All analyses were done in R with the TOSTER package (Lakens, 2017).

with our pre-registered effect size. Next, we looked at whether an inoculation effect can be detected across the two item sets within the same group (pre vs post), and found a significant difference within groups for both Group 1 from Set A (pre) to Set B (post) **[H2a]** (M_{diff} = -0.21, 95% CI_M [-0.36, -0.07], $t(238) = -3.00$, $p = .003$, $d = -0.194$, 95% CI_d [-0.322, -0.066]), as well as for Group 2 from Set B (pre) to Set A (post) **[H2b]** ($M_{diff} = -0.35$, 95% CI_M [-0.51, -0.20], $t(240) = -4.46$, $p < .001$, $d = -0.287$, 95% CI_d [-0.416, -0.158]). Finally, we looked at whether the inoculation effect changes depending on the item sets used when comparing differences-in-differences, and found no significant difference **[H3]** ($M_{diff\text{-}in\text{-}diffs} = -0.14$, 95% CI_M [-0.35, 0.07], $t(474) = -1.28$, $p = .20$, $d = -0.117$, 95% CI_d [-0.296, 0.062]). However, a TOST equivalence test could not confirm statistical equivalence to zero ($t(474) = 2.01$, p = .023).

Since for H2 and H3 different item sets were compared without standardization, we also performed the same test using *z*-scores based on the means and standard deviations of the pre-test scores for each set.[73] We found—in contrast to the non-standardized test—no significant effect for Set A-B **[H2a]** ($t(238) = -1.20$, $p = 0.23$, $d = -0.078$, 95% CI_d [-0.204, 0.049]), with statistical equivalence indicating the absence of an effect of interest ($t(238)$ = 3.44, $p < 0.001$). However, the difference between Set B-A remained significant **[H2b]** ($t(240) = -5.71$, $p < .001$, $d = -0.368$, 95% CI_d [-0.498, -0.237]). This in turn led to a significant difference-in-differences test **[H3]** ($t(476) = -3.34$, $p < .001$, $d = -0.305$, 95% CI_d [-0.485, -0.124]). See Supplementary Figure S6 for a visual plotting of the standardized scores. Figure 7.1.4.2 shows the results of the hypothesis tests.

[73] Formula used: $(Score_{SetX,T2} - M_{SetX,T1}) / SD_{SetX,T1.}$

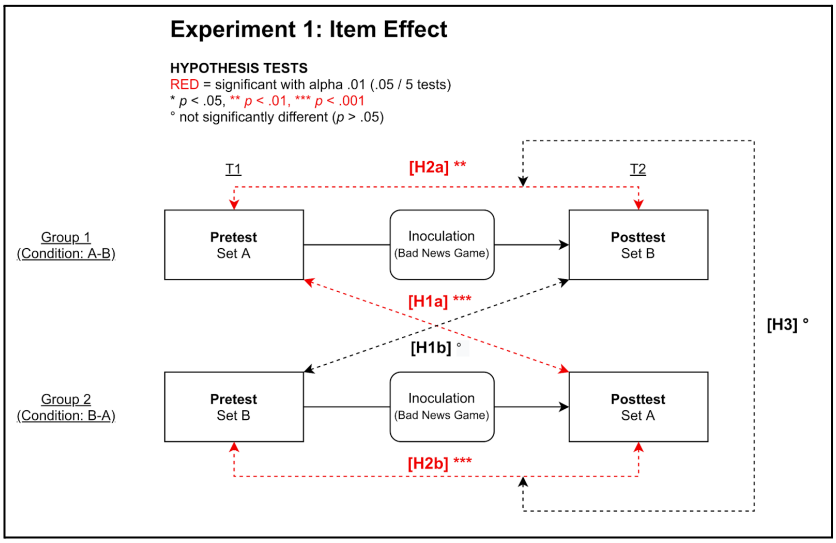

Figure 7.1.4.2. Procedure flowchart with hypothesis tests in the *Item Effects* experiment (N = 480).

We thus find partial support for our preregistered hypotheses. Our analyses indicate that the inoculation effect does not always persist in different circumstances with varying items, as the strength of the effect is influenced by the psychometric properties of the news items used and the order of the item sets. Comparing Set B (pre-test) to Set B (post-test) not only showed no significant effect, but the equivalence test also indicated absence of an effect with d = (-)0.30. This suggests that Set B did not yield the inoculation benefit across groups while Set A did. When standardizing the different item sets before comparison (to eliminate confounds with item properties), we no longer find a significant effect for Item Set A (pre-test) compared to Set B (post-test). This suggests that on top of differential effects depending on the item properties, order effects are present as well.

Real News

As preregistered, we also report our findings for the real news items. We found that, due to the low item count, the real news indices did not yield an acceptable internal consistency ($M_{real,SetA}$ = 4.50, SD = 1.32, Cronbach's α = 0.14; $M_{real,SetB}$ = 5.08, SD = 1.41, Cronbach's α = 0.43). It is therefore not a sufficient measure to come to generalizable

conclusions regarding item effects, in contrast to the fake news indices (which show an acceptable internal consistency for both item sets).

Nonetheless, using the unstandardized sets, we found no significant effect for Set A (pre-test Group 1) vs Set A (post-test Group 2) **[H1a]** (M_{diff} = -0.16, 95% CI_M [-0.39, 0.06], $t(472) = -1.43$, $p = .15$, $d = -0.131$, 95% CI_d [-0.310, 0.049]), but could not confirm statistical equivalence to zero ($t(472) = 1.85$, $p = .032$). No significant difference was found for Set B (pre-test Group 2) vs Set B (post-test Group 1) **[H1b]** (M_{diff} = 0.01, 95% CI_M [-0.26, 0.28], $t(470) = 0.05$, $p = .96$, $d = 0.005$, 95% CI_d [-0.174, 0.184]), and we were able to confirm statistical equivalence to zero ($t(470) = -3.24$, $p < 0.001$). Looking within the two item sets presented in the same group (pre vs post), we found a significant difference within groups both for both Group 1 from Set A (pre) to Set B (post) **[H2a]** (M_{diff} = 0.59, 95% CI_M [0.37, 0.80], $t(238) = 5.38$, $p < .001$, $d = 0.348$, 95% CI_d [0.217, 0.478]), and for Group 2 from Set B (pre) to Set A (post) **[H2b]** (M_{diff} = -0.74, 95% CI_M [-0.95, -0.54], $t(240) = -7.10$, $p < .001$, $d = -0.457$, 95% CI_d [-0.589, -0.324]). These effects were in the opposite direction of each other and therefore point towards item effects rather than inoculation effects (i.e., these effects can be best explained by differences in item-set baseline scores, see standardized analyses below). To see if there are differences in these item effects, we looked at changes using the difference-in-differences, and found a significant effect **[H3]** ($M_{diff\text{-}in\text{-}diffs}$ = -1.33, 95% CI_M [-1.63, -1.03], $t(477) = -8.80$, $p < .001$, $d = -0.803$, 95% CI_d [-0.995, -0.610]).

As with the fake news items, we compared the *z*-score standardized sets, and found no significant effect for **[H2a]** ($t(238) = 0.06$, $p = .95$, $d = 0.004$, 95% CI_d [-0.123, 0.131]), and TOST statistical equivalence to zero ($t(759) = -8.16$, $p < .001$). We also found no significant effect for **[H2b]** ($t(240) = -1.62$, $p = .11$, $d = -0.105$, 95% CI_d [-0.231, 0.022]), and equivalence to zero ($t(759) = 5.38$, $p < .001$). This was further corroborated by a non-significant difference-in-differences test **[H3]** $t(477) = -1.17$, $p < .241$, $d = -0.107$, 95%

CI_d [-0.286, 0.072]. See Supplementary Figure S6 for a visual plotting of the standardized scores.

These results provide evidence for all our preregistered hypotheses with respect to real news items (H1-H3). First, we found, as expected, no inoculation effect for the real news items (i.e., intervention does not increase general scepticism for news, but only for fake news) (H1a and H1b). Equivalence testing confirmed the absence of any effect of interest. Secondly, we unexpectedly found item effects when comparing the different item sets across groups (H2a and H2b), but these effects were eliminated after standardizing the scale. These findings indicate the absence of a negative inoculation effect for real news items. Raw means, standard deviations, and visuals analyses for the real news scale can be found in Supplementary Table S3 and Supplementary Figures S1 and S3.

Experiment 2: Testing Effects[74]

Fake News

In Experiment 2 we investigated the testing effect[75]. Using Solomon's Three Group analysis[76], we executed five hypothesis tests with a Bonferroni corrected $\alpha = 0.01$ (0.05/5 tests) threshold for ratings of the reliability of the various news items ($M_{\text{fake,PrePost}} = 2.73$, $SD = 1.47$, Cronbach's $\alpha = 0.81$; $M_{\text{fake,Control}} = 2.83$, $SD = 1.18$, Cronbach's $\alpha = 0.66$). First, to measure the total effect, we looked at the standard pre-test – inoculation – post-test group (further henceforth referred to as Group 1), and found a significant effect, measured by looking at the difference in mean reliability rating post-test compared to the pre-test, in line with our hypothesis **[H1]** ($M_{\text{diff}} = -0.41$, 95% CI_M [-0.53, -0.29], $t(312) = -6.30$, $p < .001$, $d = -0.356$, 95% CI_d [-0.470, -0.242]). This indicates a small total effect, but on its own does not

[74] The AsPredicted.org pre-registration can be found here:https://aspredicted.org/wy59x.pdf, all alterations are explained in Supplementary Declaration S2.

[75] We used a mixed item set (combination of Set A and Set B) for the control group; as explained above, this was an error in the implementation of the in-game survey. To control for confounds because of this, we have looked at the same analysis performed only with the shared items between all conditions, and find similar results, and therefore do not see this as a major limitation to the results presented here. See Supplementary Analysis S2 and Supplementary Figure S5 for a detailed overview.

[76] The traditional Solomon Three Group analysis can be found in Supplementary Analysis S1.

show whether this effect is found due to the success of inoculation or due to design effects. To look at potential effects solely due to pretesting, we looked at the same difference score within the control group (Group 2), and found a trivial and non-significant effect, as hypothesized **[H2]** (M_{diff} = 0.03, 95% CI_M [-0.04, 0.11], $t(759) = 0.89$, $p = .37$, $d = 0.032$, 95% CI_d [-0.039, 0.103]) that is statistically equivalent to zero ($t(759) = -7.38$, $p < .001$). See Figure 7.1.4.3 for a bar graph (Panel A) and density plots (Panel B).

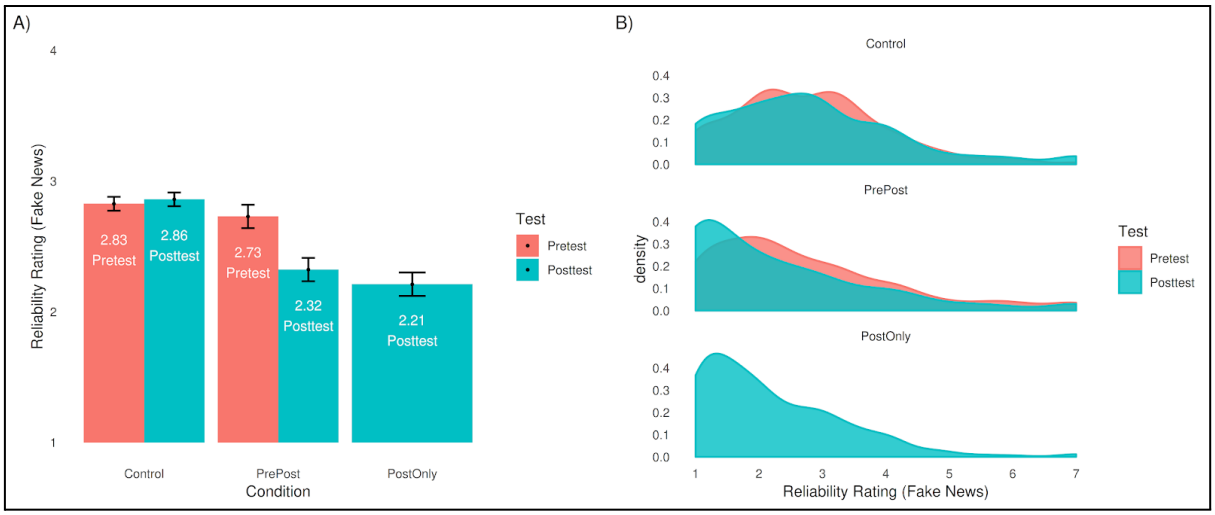

Figure 7.1.4.3. Bar chart (A) and density plots (B) of fake news reliability ratings in the *Testing Effects* study. N = 1,679. Error bars represent 95% Confidence Intervals.

Next, to investigate the interaction effect between the pre-test and the intervention, we looked at whether the post-test mean of Group 1, subtracted by the pre-test effect is different from the post-test mean of the post-test-only group (Group 3). As hypothesized, we found a positive but non-significant interaction effect **[H3]** (M_{diff} = 0.08, 95% CI_M [-0.11, 0.27], $t(503)$ = 0.82, $p = .41$, $d = 0.060$, 95% CI_d [-0.077, 0.196]), statistically equivalent ($t(502) = -3.31$, $p < .001$) Finally, we subtracted the pre-test effect and interaction effect from the post-test mean in Group 1, and then looked at the pre-post difference score. We found a significant inoculation effect **[H4]** (M_{diff} = -0.52, 95% CI_M [-0.65, -0.39], $t(312) = -8.05$, $p < .001$, d = -0.455, 95% CI_d [-0.571, -0.338]), in line with our hypothesis. Finally, to confirm our

Solomon analysis in an alternative test, which is feasible because of the high sample size, we looked at the difference between the post-test mean in Group 3 in comparison to the pre-test mean in Group 1, and found a significant effect similar to the one found with the traditional analysis (as hypothesized) **[H5]** (M_{diff} = -0.52, 95% CI_M [-0.71, -0.33], $t(506) = -5.47$, $p < .001$, d = -0.396, 95% CI_d = [-0.534, -0.258]). These hypothesis tests indicate that the inoculation effect is not affected by the administration of a pre-test or a pre-test X intervention interaction, and that even if there would be a testing effect, it is a small one that does not amplify but slightly *decreases* the inoculation effect. See Figure 7.1.4.4 for a flowchart with indicated results of our hypothesis tests.

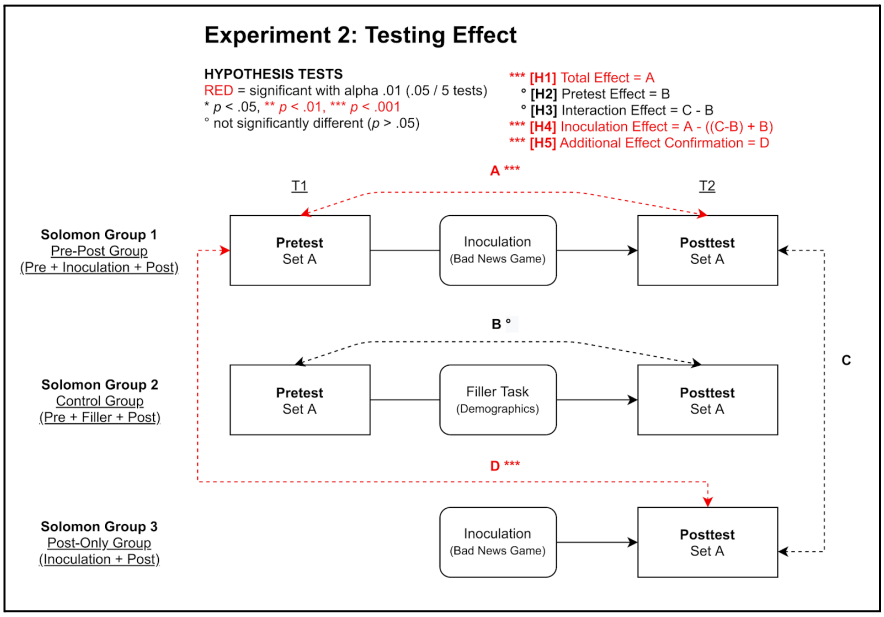

Figure 7.1.4.4. Procedure flowchart with hypothesis tests in the *Testing Effects* experiment (N = 1,679).

Real News

For Experiment 2, we again found low internal consistency for the real news items ($M_{real,PrePost}$ = 4.18, SD = 1.57, Cronbach's α = 0.39; $M_{real,Control}$ = 5.17, SD = 1.58, Cronbach's α = 0.43). As in Experiment 1, this is most likely due to the low number of real news items (2)

used in our study, as the fake news items (6) do show acceptable internal consistency. Therefore the two real news items are not a sufficient measure to come to generalizable conclusions regarding testing effects with regards to real news.

Nonetheless, we first looked at the overall effect, and found a negative effect **[H1]** (M_{diff} = -0.39, 95% CI_M [-0.58, -0.20], $t(312) = -3.95$, $p < .001$, d = -0.223, 95% CI_d [-0.335, -0.111]), indicating an unexpected decrease in reliability rating for real news items, not in line with our hypothesis. We found no significant pre-test effect that could explain this **[H2]** (M_{diff} = -0.10, 95% CI_M [-0.20, -0.01], $t(759) = -2.09$, $p = .04$, d = -0.076, 95% CI_d [-0.147, -0.005]), and were able to confirm statistical equivalence to zero ($t(759) = 6.18$, $p < .001$). We did find an interaction effect between the pre-test and the intervention **[H3]** (M_{diff} = -0.31, 95% CI_M [-0.55, -0.08], $t(522) = -2.62$, $p = .009$, d = -0.188, 95% CI_d [-0.325, -0.051]). We did not find a pure inoculation effect (i.e., the inoculation effect when subtracting pre-testing effects and item set X intervention interaction effects) for real news **[H4]** (M_{diff} = 0.03, 95% CI_M [-0.17, 0.22], $t(312) = 0.27$, $p = .79$, d = 0.015, 95% CI_d [-0.096, 0.126]). Statistical equivalence confirmed to zero ($t(609) = -3.56$, $p < .001$), indicating that the observed decreased reliability rating was due to the interaction and not due to a generalized negative influence of the intervention. Our final hypothesis test, which compared the posttest in the Intervention-Post group to the pre-test of the Pre-Intervention-Post group, further corroborated evidence for the absence of a negative inoculation effect **[H5]** (M_{diff} = 0.03, 95% CI_M [-0.18, 0.24], $t(596) = 0.25$, $p = .80$, d = 0.018, 95% CI_d [-0.119, 0.154]), with statistical equivalence equal to zero ($t(596) = -4.02$, $p < .001$).

The overall effect indicated a negative impact of the intervention, which on the surface could be interpreted as an increase in scepticism towards real as well as fake news. However, the pure inoculation effect (total effect minus the pre-test and pretest-intervention interaction effects) was not only insignificant, but any effect of interest was absent. As a

significant interaction was found between pre-test and intervention (a pre-test effect was absent), these results suggest that while negative effects can be expected when using the intervention with real news items in pre-post designs, these effects are due to an interaction with the specific item set and not due to a negative inoculation effect of the intervention itself. We can thus confirm our hypotheses for the real news items. Raw means, standard deviations, and visuals analyses for the real news items can be found in Supplementary Table S5, and Supplementary Figures S4 and S5.

7.1.5 Discussion

"The common procedure has been to give a group a pre-test, using an acceptable attitude scale, then subject the group to educational procedures of some sort, and then post-test the group with the same test or an equivalent form of it." – Richard Solomon (1949, p. 139)

The procedure outlined by Solomon in 1949 is still used today, for example, in evaluating fake news interventions (Roozenbeek & van der Linden, 2019a). Because large within-designs allow for greater measurement precision than noisy small sample between-designs, Gelman (2017) suggests that psychologists should routinely use them. Yet, do within-designs "poison the well"? This study has addressed two major open methodological questions about the effectiveness of inoculation interventions in the context of online misinformation. First, we found that using different testing item sets for pre- and post-intervention scores we do not always find significant reductions with our preregistered Smallest Effect Size of Interest (SESOI) of d = -0.30. We found two types of item effects to be present: *item set order effects* and *differences in psychometric properties* between the item sets. Comparing Set B (pre-test) to Set B (post-test) across groups, for example, did not yield any effect of interest, indicating a potential validity problem. Comparing item Set A to item Set B yielded different results than comparing Set B to Set A, indicating order effects. This finding may be explained by an overestimation on our part of the preregistered effect size,

and basing our power and SESOI on a more potent item set. We refer here to Supplementary Table S4: comparing Set A (pre) to Set B (post) gives an effect size of d = -0.39 (CI [-0.45, -0.33]), which is higher than the preregistered d = 0.30, whereas comparing Set B (pre) to Set B (post) gives d = -0.23 (CI [-0.29, -0.17]). This indicates that the effect size differs from set to set, and that the preregistered value of d = 0.30 was too optimistic for Set B. Nonetheless, d = 0.23 (which would have yielded no statistical equivalence to zero) is considered a meaningful effect size for interventions in persuasion research (Funder & Ozer, 2019). To summarize, we find that inoculation still occurs when using different items in the pre- and post-test. However, when doing so, we do find fairly minimal item effects.

The raw differences between pre- and post-scores, moreover, were consistently in the right direction (see Figure 7.1.4.1): participants systematically gave lower average reliability ratings to fake news after the intervention. In the case of real news, we only found differences in starting values. After standardizing the sets to be able to compare them without the confound of differential item properties, differences in the perceived reliability of real news before and after playing the *Bad News* game disappeared. This points towards the importance of item design and compatibility of the different real news headlines.

Combined, these results suggest that people who play the *Bad News* game indeed show improvement in their latent ability to spot misinformation *techniques*, as opposed to merely improving their ability to do so in examples of misinformation that they had already seen before (in the pre-test). However, the experiment also shows that the sets of items used and the order in which they are presented can make a difference for the effect size. Future research as well as practitioners should take this into account. One way to do this is to aim towards the development of better psychometrically validated and equalized item sets that combine real and fake news items that ideally could be used for a wide range of misinformation interventions.

Second, using Solomon's Three Group Design, we demonstrate that testing effects in inoculation interventions are minimal. If there is any effect at all, it is small and in the opposite direction than expected: participants who do *not* go through a pre-test (Group 3) perform *better* at spotting misinformation and discerning real news from fake news after playing *Bad News* than participants who take both a pre-test and a post-test (Group 1). The Solomon analysis suggests that if there is an effect, this would be due to interaction effects between the pre-test and the intervention. That is, there is something about the way the pre-test interacts with the intervention that ultimately reduces its effectiveness. This suggests that researchers and practitioners can consider omitting the pre-test for large-sample mixed-design (pre-test vs. post-test and control vs. treatment) studies on inoculation against online misinformation without having to worry about detrimental consequences for the effectiveness of the intervention.

With respect to real news items, we found a significant interaction effect between the item set and the intervention, which causes a negative overall effect. This can give the impression that the *Bad News* intervention has negative side effects by reducing trust in real news. However, this effect can be fully accounted for by the interaction between the specific item set and the intervention. Moreover, the pure inoculation effect was insignificant and showed the absence of any effect of interest using a TOST equivalence test with Cohen's d = (-)0.30 as the SESOI. Since in Experiment 1 we did not find any effect for real news items either, we conclude that the *Bad News* game does not increase scepticism towards real news. However, we caution to give too much weight to this result, as the limited internal consistency of the real news items in combination with the potential item-specific effects makes it difficult to draw general conclusions.

One of the principle tenets proposed by Solomon (1949) was that given a large enough, randomly selected sample size, with known means and variances of a pre-test for two

groups (i.e., Group 1 and 2), one can infer what the pre-test result would be for a third group (i.e., Group 3). The inclusion of the traditional control group (i.e., Group 2, which only has a pre and post-test without an intervention) presumably serves two purposes. First, at the time of Solomon's writing, sample sizes of hundreds or even thousands of participants (which are achievable today through online interventions), were arguably unthinkable. As such, by averaging the pre-test score of Group 2 with the pre-test score of Group 1, researchers could reasonably assume a third pre-test for Group 3 with a much larger sample size. This would allow for researchers to determine the traditional within (i.e., pre versus post-test) analysis for Group 3. Second, the inclusion of Group 2 allows for the assessment of the unique influence of a pre-test as well as the pre-test X intervention on the post-test. For the former, in a tightly controlled, short duration, laboratory or online study, the difference between pre-test and post-test without an intervention is likely to be minimal, if not zero (as was the case in this study). But more importantly, as is usually the case in psychological research, researchers are more concerned with the true influence of the intervention on the post-test score (i.e., Group 3), which can be isolated without the second group. We therefore argue, in agreement with Solomon's original tenet, that if the primary concern is to determine the influence of an intervention on the post-test (when there is no pre-test), a first control group is not necessary (i.e., Group 2)[77]. Instead, one would only need to include a group that went through the pre-test, intervention, and post-test, and a group that only went through the intervention and post-test. As an example, had we only included Group 1 and Group 3 in this study, we would have observed a pretest X intervention interaction effect (which now includes the pre-testing effect) leading to an inoculation effect mean difference of 0.11 instead of 0.08 (which excludes the pre-testing effect), and the same corrected inoculation effect of a -0.52 decrease in reliability ratings of fake news after the *Bad News* game intervention.

[77] When the following conditions are met: a sufficiently large sample size, random selection, and known variance and mean of one group (Solomon, 1949).

For illustrative purposes, we used a concrete example of how careful attention to design effects can enhance the quality of psychological research on important real-world issues such as fake news, but we believe that the method and results outlined in this paper are relevant to many psychological interventions in educational settings, from reducing prejudice and stereotypes, to work on memory and reaction times.

We do, however, note a number of limitations. In the item selection process for the item effects study, we did not correctly register the polarization badge for the majority of the sample[78], which led to reduced power for the item effect study compared to the pre-registration (a post-hoc power analysis yields a power of .76 to detect effect sizes of d = -0.30; for comparison, post-hoc power is 0.96 for the same effect size for the testing effect study). However, in the two prior attempts at running this study (N_1 = 2,408, N_2 = 1,532) we found similar effects for each comparison, except for the Set B - Set B cross-group comparison. In both of these studies we found a significant Set B - Set B effect with mean differences of -0.35 (t(1477) = -7.31, p < .001, d = -0.321) and -0.19 (t(1028) = -3.24, p = .001, d = -0.175). We thus have reason to see the results from the main study as valid, despite the reduction in power. See Supplementary Table S4 for a comparison of the hypothesis tests between the large dataset (without polarization category) and the small dataset (with polarization category). For the testing effects study, we used a mixed item set for the control group (Group 2). Both of these limitations occurred due to a selection error in the experiment. We corrected for this using multiple rigorous and transparent analyses (see the Supplementary Analyses S1 and S2, as well as Supplementary Materials S1), which show that the conclusions presented here are nonetheless robust.

In conclusion, game-based inoculation interventions are minimally influenced by item effects, and not by testing effects. The inoculation effect generalizes across different item

[78] The first attempt did not include the polarization items, and the second included the same polarization item in both surveys, instead of two different items. Because of this, we decided not to include these datasets in our main analysis.

sets, with item effects being limited and small likely due to the use of sets that were not psychometrically standardized a priori. Pre-test – post-test designs (both within-subject and between-subject) can be used, as well as between-subject post-test-only designs, to measure inoculation effects in the context of online misinformation. Only small descriptive testing effects were found, and our investigation suggests that these potential effects are rooted in an interaction between the specific item set and the intervention and are not due to a differential intervention effect on the latent ability (i.e., the general ability to spot misinformation techniques). Future research could help inform the extent to which these findings are generalizable to other psychological interventions.

Study 7

7.2 The Misinformation Susceptibility Test (MIST): A Psychometrically Validated Measure of News Veracity Discernment[79]

7.2.1 Abstract

Interest in the psychology of misinformation has exploded in recent years. Despite ample research, to date there is no validated framework to measure misinformation susceptibility. Therefore, we introduce ***Veri**fication **do**n**e***, a nuanced interpretation schema that simultaneously considers *Veracity discernment*, and the distinct, measurable abilities (***r**eal news detection*, ***f**ake news detection*), and biases (***d**istrust*—negative judgement bias; ***n**aïvité*—positive judgement bias) that it is composed of—thus offering a nuanced assessment. We then conduct three studies with six independent samples (N_{total} = 7,291) to develop, validate, and apply the Misinformation Susceptibility Test (MIST), the first psychometrically-validated measure of misinformation susceptibility. In Study 7A (N = 409) we use a neural network language model to generate items, and use factor analysis and item-response theory to create the MIST-20 (20 items; <2 minutes) and MIST-8 (8 items; <1 minute). In Study 7B (N = 6,461) we confirm model fit in four representative samples (US, UK), from three different sampling platforms—Respondi, CloudResearch, and Prolific. We also explore the MIST's nomological net, which demonstrates good convergent and discriminant validity, and generates age-, region-, and country-specific norm tables. In Study 7C (N = 421) we demonstrate how the MIST—in conjunction with ***Veri**fication **do**n**e***—can provide novel insights on existing psychological interventions, thereby advancing theory

[79] Study 7 is currently under review at a scientific journal as *"The Misinformation Susceptibility Test (MIST): A Psychometrically Validated Measure of News Veracity Discernment"*, and has been published on PsyArXiv (Maertens et al., 2022). It is the result of a collaboration with Professor Friedrich Götz (University of British Columbia), Dr Claudia Schneider (University of Cambridge), Dr Jon Roozenbeek (University of Cambridge), Dr John Kerr (University of Cambridge), Professor Stefan Stieger (Karl Landsteiner University of Health Sciences), Dr William McClanahan (Max Planck Institute for the Study of Crime, Security, and Law), Karly Drabot (University of Cambridge), and Professor Sander van der Linden (University of Cambridge). I am joint first author on the paper together with Assistant Professor Friedrich Götz.

development. Finally, we outline the versatile implementations of the MIST as a screening tool, covariate, and intervention evaluation framework. Introducing the MIST hence not only advances misinformation scholarship, but also provides a blueprint for integrated theory and measurement development.

7.2.2 Introduction

The global spread of misinformation has had a palpable negative impact on society. Recent research showed us how misleadinging information is linked to radicalization, terrorism, propaganda, and extremism (Chiluwa, 2019; Garry et al., 2021; Piazza, 2022; Zihiri et al., 2022). Meanwhile, conspiracy theories about coronavirus disease 2019 (COVID-19) and the vaccines against it have been linked to increased vaccine hesitancy and the vandalization of cell phone masts (Hotez et al., 2021; Jolley & Paterson, 2020; Loomba et al., 2021; Roozenbeek et al., 2020). With false and moral-emotional media spreading faster and deeper than more accurate and nuanced content (Brady et al., 2017; Vosoughi et al., 2018), the importance of information veracity has become a central debate for scholars and policy-makers (Lewandowsky et al., 2017, 2020).

Accordingly, across disciplines, research on the processes behind, impact of, and interventions against misinformation—which has been around for decades—has surged over the past years (for recent reviews, see Pennycook & Rand, 2021; Van Bavel et al., 2021; van der Linden et al., 2021). Researchers have made progress in designing media and information literacy interventions in the form of educational games (Basol et al., 2021; Roozenbeek & van der Linden, 2019a, 2020), "accuracy" primes (Pennycook, Epstein, et al., 2021; Pennycook, McPhetres, et al., 2020), introducing friction (Fazio, 2020), and inoculation messages (Lewandowsky & van der Linden, 2021). Crucially, however, no theoretical framework exists for a nuanced evaluation of misinformation susceptibility, nor a psychometrically validated measurement that provides a reliable measure across studies.

Inconsistent Interpretation and the Need for a New Measurement Instrument

Despite the plethora of research papers on the psychology of misinformation, the field has not converged on a standardized way of defining or measuring people's susceptibility to misinformation. In the absence of such a commonly agreed-upon standard, scholars have been inventive in the way that they employ individually constructed misinformation tests, often with the best intentions to create a good scale, but typically without formal validation (e.g., Pennycook, Epstein, et al., 2021; Roozenbeek, Maertens, et al., 2021).

The extent of the problem becomes evident when examining how researchers develop their test items and report the success of their models or interventions. Typically, researchers create (based on commonly used misinformation techniques; e.g., Maertens et al., 2021; Roozenbeek & van der Linden, 2019a) or select (from a reliable fact-check database; e.g., Cook et al., 2017; Guess et al., 2020; Pennycook, McPhetres, et al., 2020; Pennycook & Rand, 2019; Swire, Berinsky, et al., 2017; van der Linden et al., 2017) news headlines or social media posts, where participants rate the reliability, sharing intention, accuracy, or manipulativeness of these items on a Likert or binary (e.g., true vs. false) scale. Sometimes the news items are presented as plain-text statements (e.g., Roozenbeek et al., 2020), while in other studies researchers present headlines together with an image, source, and lede sentence (e.g., Pennycook & Rand, 2019). The true-to-false ratio often differs, where in some studies only false news items are presented (e.g., Roozenbeek et al., 2020), in others this is an unbalanced (e.g., Roozenbeek, Maertens, et al., 2021) or balanced (e.g., Pennycook & Rand, 2019) ratio of true and false items. Often an index score is created by taking the average of all item ratings (an index score reflecting general belief in false or true news items; e.g., Maertens et al., 2021), or by calculating the difference between ratings of true items and false items (veracity discernment; e.g., Pennycook, McPhetres, et al., 2020). Finally, an effect size is calculated, and a claim is made with respect to the effectiveness of the intervention, based

on a change in false news ratings (e.g., Roozenbeek & van der Linden, 2019a), a combined change in true news ratings and false news ratings (e.g., Guess et al., 2020), or even a change in true news ratings only (Pennycook, McPhetres et al., 2020).

It becomes clear that the wide variation in methodologies makes it hard to compare studies or generalize conclusions beyond the studies themselves. Little is known about the psychometric properties of these ad hoc scales and whether or not they measure a latent trait. We often assume they do, but we do not know if they are measuring the same construct, and could be engaging in a common essence bias (Brick et al., 2021). We currently do not know how different scales are related, or how the true-to-false ratios influence their outcome (Aird et al., 2018), and how much of the effects found are due to response biases rather than changes in skill (Batailler et al., 2020). The limited studies that do look at the issue of scale-specific effects, show significant item effects, indicating a risk of skewed conclusions about intervention effect sizes (e.g., Roozenbeek, Maertens, et al., 2021). Relatedly, whether the sampling of test items, their presentation, and response modes have a high ecological validity is often not discussed (Dhami et al., 2004), and little is known about the nomological net and reliability of the indices used. In other words, it is difficult to disentangle whether differences between studies are due to differences in the interpretation schema, the measurement instrument, or actual differences in a misinformation susceptibility. This indicates a clear need for a unified theoretical framework in conjunction with a standardized instrument with strong internal and external validity.

Towards A Universal Conceptualization and Measurement: The Verification done Framework

Here, we set out to create a theoretical interpretation schema as well as a first psychometrically validated measurement instrument that, in conjunction, resolve the issues mentioned above and has utility for a wide range of scholars. We extend the current literature by providing the first framework and measurement instrument that allows for a reliable *holistic* measurement through the ***Ver**i**f**ication **d**o**n**e* framework: we can only fully interpret misinformation susceptibility or the impact of an intervention by capturing *news veracity discernment* (**V**, ability to accurately distinguish real news from fake news) as a general factor, the specific facets *real news detection ability* (**r**, ability to correctly identify real news) and *fake news detection ability* (**f**, ability to correctly identify fake news), *distrust* (**d**; negative judgement bias / being overly skeptical) and *naïvité* (**n**; positive judgement bias, being overly gullible), and comparing **V**, **r**, **f**, **d**, *and* **n** alongside each other. For example, two different interventions may increase discernment ability **V** to a similar extent, but intervention A might do so by increasing detection ability **r**, while intervention B may accomplish the same by increasing **f**. Similarly, two people with the same discernment ability **V** may have opposite **r** and **f** abilities. Changes in detection abilities **r** or **f** after an intervention have to be interpreted together with changes in judgement biases **d** and **n** to figure out whether the intervention has done more than just increase a judgement bias. Existing interventions often look at a limited subset of these five dimensions, for example the creators of the Bad News Game intervention (Roozenbeek & van der Linden, 2019a) originally focused on *fake news detection*, only including a few real news items. Meanwhile, the *accuracy nudge* intervention seems to mainly work by addressing *real news detection* (Pennycook, McPhetres, et al. 2020), although we are not sure about the judgement biases.

Another media literacy intervention was found to increase general distrust, but in general, showed improvement on *veracity discernment* nevertheless (Guess et al., 2020).

In order to be able to compare these scores and gain insights into the complete picture, we need to employ the ***Verification done*** framework, but also make sure that each scale has a high validity and comparability. To accomplish this, through a series of three studies and using a novel neural-network-based item generation approach, we develop the Misinformation Susceptibility Test (MIST): a *psychometrically validated* (i.e., based on classical test theory and item-response theory) measurement instrument. The MIST was developed to be the first truly balanced measure with an equal emphasis on discernment, real news detection, fake news detection, and judgement bias. In addition, to put the results into perspectives, all scores should be interpreted along with nationally representative norm tables. In the present study, we describe how we developed and validated the MIST to accomplish these goals, evaluate each of these dimensions, and investigate the practical utility of the MIST for researchers and practitioners in the field.

The Misinformation Susceptibility Test

We conduct three studies to develop, validate, and apply the MIST. In Study 7A (N = 409), employing a multitude of exploratory factor analysis (EFA)- and item response theory (IRT)-based selection criteria to create a 20-item MIST full-scale and an 8-item MIST short-scale from a larger item pool that had first been created by a combination of advanced language-based neural network algorithms and real news headline extraction from reliable and unbiased media outlets and then been pre-filtered through multiple iterations of expert review. The resultant MIST scales are balanced (50% real, 50% fake), binary, cumulatively scored instruments that ask participant to rate presented news headlines as either true or false, with higher MIST scores indicating greater media literacy. As such, the MIST exhibits a

higher-order structure, with two first-order factors (i.e., real news detection, fake news detection) and one general ability second-order factor (i.e., veracity discernment).

In Study 7B (N = 6,461), we employ CFAs to replicate the MIST's structure across four nationally representative samples from the UK and the US, establish construct validity via a large, preregistered nomological network and derive norm tables for the general populations of the UK and US, and demographic and geographic subgroups.

In Study 7C (N = 421), we provide an example of how to implement ***Verification done*** and the MIST in the field by applying it in the naturalistic setting of a well-replicated media literacy intervention, the *Bad News Game*. Whereas ample prior studies have attested to the theoretical mechanisms and effects that contribute to the *Bad News Game*'s effectiveness in reducing misinformation susceptibility (see e.g., Maertens et al., 2021; Roozenbeek & van der Linden, 2019a), within-subject repeated-measures analyses of the MIST-8 for pre-and post-game tests in conjunction with the ***Verification done*** framework reveal important new insights about how the intervention affects people across different evaluative dimensions. This paper demonstrates the benefits of integrated theory and assessment development, resulting in a framework providing nuanced, multi-faceted insights that can be gained from a short, versatile, psychometrically sound and easy-to-administer new measure. Table 7.2.2.1 offers a comprehensive summary of all samples used, detailing their size, demographic breakdowns, included measures, country of origin, recruitment platform and whether or not they were nationally representative and preregistered.

Table 7.2.2.1
Summary of Samples

	Study 7A: Development	Study 7B: Validation				Study 7C: Application
Sample	1	2A	2B	2C	2D	3
N	409	3,479	510	1,227	1,245	421
Country of Origin	USA	USA	USA	UK	UK	USA
Nationally Representative Quota	No	Yes	Yes	Yes	Yes	No
Recruitment Platform	Prolific	Respondi	CloudResearch	Respondi	Prolific	Bad News Game
Preregistration	Yes	No	Yes	No	No	No
Demographic Composition	**Age** M_{age} = 33.20 SD_{age} = 11.85 **Gender** 55.50% female 42.30% male 2.20% other / non-binary **Ethnicity** /	**Age** M_{age} = 45.10 SD_{age} = 16.16 **Gender** 51.11% female 48.84% male 0.06% other / non-binary **Ethnicity** 76.89% White, Caucasian, Anglo, or European American 8.39% Asian or Asian American 6.00% Hispanic or Latino 5.98% Black or African American 1.12% Native American or Alaskan Native 0.54% Middle Eastern 0.30% Hawaiian or Pacific Islander 0.77% Other/Prefer not to answer	**Age** M_{age} = 49.25 SD_{age} = 16.96 **Gender** 55.88% female 43.53% male 0.59% other / non-binary **Ethnicity** 68.81% White, Caucasian, Anglo, or European American 4.28% Asian or Asian American 11.05% Hispanic or Latino 12.12% Black or African American 2.50% Native American or Alaskan Native 0.18% Middle Eastern 1.07% Other/Prefer not to answer	**Age** M_{age} = 45.34 SD_{age} = 16.52 **Gender** 51.67% female 48.33% male 0.00% other / non-binary **Ethnicity** 87.33% White 6.95% Asian 2.45% Black 0.08% Arab 2.13% Mixed 1.06% Other	**Age** M_{age} = 44.66 SD_{age} = 15.65 **Gender** 52.53% female 47.07% male 0.40% other / non-binary **Ethnicity** 86.10% White 7.47% Asian 3.53% Black 0.16% Arab 1.61% Mixed 1.12% Other	**Age** 55.58% [18, 29] 32.30% [30, 49] 12.11% [50, 99] **Gender** 52.02% female 41.09% male 6.89% other / non-binary **Ethnicity** /

	Education 1.47% Less than high school degree 9.29% High school graduate 31.30% Some college but no degree 38.88% Bachelor's degree in college 1.96% Professional degree 13.45% Master's degree 3.67% Doctoral degree	**Education** 1.74% Did not complete high school 34.98% High school degree or equivalent 15.08% Associate's degree 31.84% Degree (Bachelors) or equivalent 15.11% Degree (Masters) or other postgraduate qualification 1.25% Doctorate 0.97% Other/Prefer not to say	**Education** 2.55% Less than high school degree 25.10% High school graduate 27.45% Some college but no degree 26.08% Bachelor's degree in college 1.57% Professional degree 13.92% Master's degree 3.33% Doctoral degree	**Education** 11.03% No formal education above age 16 16.18% Professional or technical qualifications above age 16 27.12% School education up to age 18 31.94% Degree (Bachelors) or equivalent 12.09% Degree (Masters) or other postgraduate qualification 1.63% Doctorate	**Education** 6.27% No formal education above age 16 10.68% Professional or technical qualifications above age 16 25.22% School education up to age 18 38.63% Degree (Bachelors) or equivalent 16.87% Degree (Masters) or other postgraduate qualification 2.33% Doctorate	**Education** 14.49% High school or less 36.10% Some college 49.41% Higher degree
Measured Constructs	*- MIST-8/MIST-20* *- BSR* *- CMQ* *- COVID-19 compliance* *- CRT* *- DEPICT* *- CV19 fact-check*	*- MIST-8/MIST-20* *- MIST-8* *- AOT* *- Anti-vaccination attitudes* *- COVID-19 misinformation beliefs* *-CRT* *- Numeracy* *- Political ideology* *- Trust (in scientists, journalists, politicians, the government)*	*- MIST8/MIST-20* *- BSR* *- BFI2-S* *- CMQ* *- EDO* *- DEPICT SF* *- Go Viral!* *- MFQ-S* *- SD4* *- SDO* *- SINS* *- SISES* *- SIRIS* *- SSPC* *- Trust (in medical personnel, scientists, politicians, journalists, the government, scientific knowledge, civil servants, mainstream media)*	*- MIST-8/MIST-20* *- Political ideology* *- Trust (in scientists, journalists, politicians, the government)*	*- MIST-8/MIST-20* *- Political ideology* *- Trust (in scientists, journalists, politicians, the government)*	*- MIST-8* *- BN*

Note. AOT = Actively Open-minded Thinking (Baron, 2019); BFI-2-S = Big-Five Inventory 2 Short-Form (Soto & John, 2017); BN = Bad News Game (Roozenbeek & van der Linden, 2019a); BSR = Bullshit Receptivity scale (Pennycook et al., 2015); CMQ = Conspiracy Mentality Questionnaire (Bruder et al., 2013); CRT = Cognitive Reflection Test (Frederick, 2005); DEPICT = Discrediting-Emotion-Polarization-Impersonation-Conspiracy-Trolling deceptive headlines inventory (Maertens et al., 2021); DEPICT SF = DEPICT Balanced Short Form (Maertens et al., 2021); EDO = Ecological Dominance Orientation (Uenal et al., 2022); CV19 fact-check = COVID-19 fact-check test (Pennycook, McPhetres, et al., 2020); Go Viral! = Go Viral! Balanced Item Set (Basol et al., 2021); MFQ-S = Moral Foundations Questionnaire Short Version (Graham et al., 2011); Numeracy = combination of Schwartz Numeracy Test (Schwartz et al., 1997) and Berlin Numeracy Test (Cokely et al., 2012), SD4 = Short Dark Tetrad (Paulhus et al., 2020); SDO = Social Dominance Orientation (Ho et al., 2015); SINS = the Single-Item Narcissism Scale (Konrath et al., 2014); SISES = Single-Item Self-Esteem Scale (Robins et al., 2001) SIRIS = Single-Item Religious Identification Scale (Norenzayan & Hansen, 2006); SSPC = Short Scale of Political Cynicism (Aichholzer & Kritzinger, 2016).

Study 7A

7.3 Development—Scale Construction and Exploratory Psychometric Analyses

7.3.1 Methods

Following classic (Clark & Watson, 1995; Loevinger, 1957) and recent (Boateng et al., 2018; Rosellini & Brown, 2021; Zickar, 2020) psychometrics guidelines, and taking into account insights from misinformation scholars (Pennycook, Binnendyk, et al., 2021; Roozenbeek, Maertens, et al., 2021), we devised a four-stage scale development protocol (i.e., 1—item generation, 2—expert filtering, 3—quality control, and 4—data-driven selection), shown in Figure 7.3.1.1.

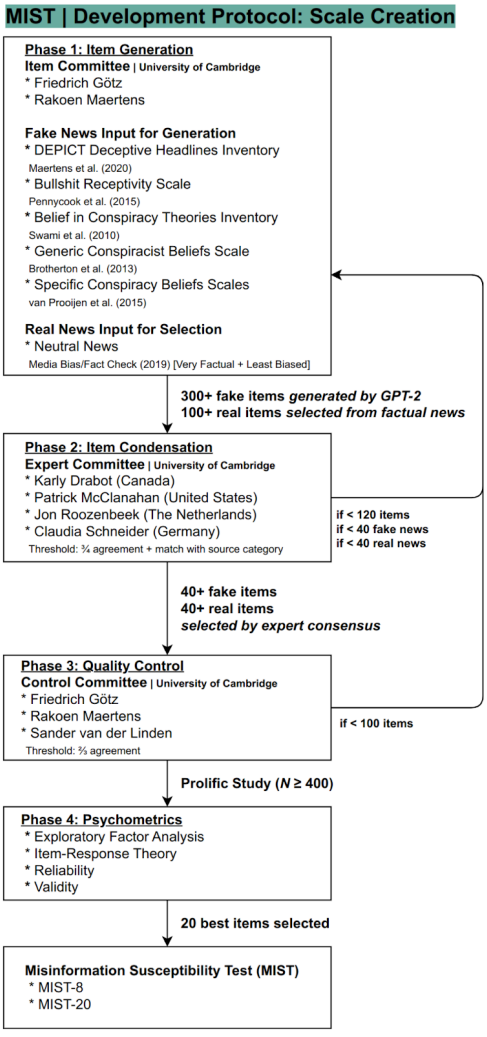

Figure 7.3.1.1. Development protocol of the Misinformation Susceptibility Test.

Preparatory Steps

Phase 1: Item Generation.

Fake News. There is a debate in the literature on whether the misinformation items administered in misinformation studies should be actual news items circulating in society, or news items created by experts that are fictional but feature common misinformation techniques. The former approach arguably provides better ecological validity (Pennycook, Binnendyk, et al., 2021), while the latter provides a cleaner and less confounded measure since it is less influenced by memory and identity effects (van der Linden & Roozenbeek, 2020). Considering these two approaches and reflecting on representative stimulus sampling (Dhami et al., 2004), we opted for a novel approach that combines the best of both worlds. We employed the generative pre-trained transformer 2 (GPT-2)—a neutral-network-based artificial intelligence developed by OpenAI (Radford et al., 2019)—to generate fake news items. The GPT-2 is one of the most powerful open-source text generation tools currently available for free use by researchers (Götz, Maertens, et al., 2021). It was trained on eight million text pages, combines 1.5 billion parameters, and is able to write coherent and credible articles based on just one or a few words of input.[80] We did this by asking the GPT-2 to generate a list of fake news items inspired by a smaller set of items. This smaller set contained items from any of five different scales that encompass a wide range of misinformation properties: the Belief in Conspiracy Theories Inventory (BCTI; Swami et al., 2010), the Generic Conspiracist Beliefs scale (GCB; Brotherton et al., 2013), specific Conspiracy Beliefs scales (van Prooijen et al., 2015), the Bullshit Receptivity scale (BSR; Pennycook et al., 2015), and the Discrediting-Emotion-Polarization-Impersonation-Conspiracy-Trolling deceptive headlines inventory (DEPICT; Maertenset al., 2021; Roozenbeek & van der Linden, 2019a). We set out

[80] For a step-by-step guide on how to set up the GPT-2 to use as a psychometric item generator, see the tutorial paper by Götz, Maertens, et al. (2021), as well as the useful blog posts by Woolf (2019), Nasser (2020), and Curley (2020).

to generate 100 items of good quality but as this is a new approach, we opted for the generation of at least 300 items. More specifically, we let GPT-2 generate thousands of fake news headlines, and cleaned out any duplicates and clearly irrelevant items (see Supplement S1 for a full overview of all items generated and those that have been removed).

Real News. For the real news items, we decided to include items that met each of the following three selection criteria: (1) the news items are actual news items (i.e., they circulated as real news), (2) the news source is the most factually correct (i.e., accurate), and (3) is the least biased (i.e., non-partisan or politically centrist). To do this, we used the Media Bias Fact Check (MBFC; https://mediabiasfactcheck.com/) database to select news sources marked as *least biased* and scoring *very high* on factual reporting.[81] The news sources we chose were Pew Research (https://www.pewresearch.org/), Gallup (https://www.gallup.com/), MapLight (https://maplight.org/), Associated Press (https://www.ap.org/), and World Press Review (http://worldpress.org/). We also diversified the selection by including the non-US outlets Reuters (https://www.reuters.com/), Africa Check (https://africacheck.org/), and JStor Daily (https://daily.jstor.org/). All outlets received the maximum MBFC score at the time of item selection.[82] A full list of the real news items selected can be found in Supplement S1.

Overall, this item-generation process resulted in an initial pool of 413 items. The full list of items we produced and through which methods each of them was attained can be found in Supplement S1.

Phase 2: Item Condensation. To reduce the number of headlines generated in Phase 1, we followed previous scale development research and practices (Carpenter, 2018; Haynes et al., 1995; Simms, 2008) and established an Expert Committee with misinformation

[81] MBFC is an independent fact-checking platform that rates media sources on factual reliability as well as ideological bias. At the time of writing, the MBFC database lists over 3,700 media outlets and its classifications are frequently used in scientific research (e.g., Bovet & Makse, 2019; Chołoniewski et al., 2020; Cinelli et al., 2021).

[82] Three out of six no longer receive the maximum score, and are now considered to have a *center-left* bias, and score between *mostly factual* and *highly factual* reporting: World Press Review (*mostly factual*, *center-left*), MapLight (*highly factual*, *center-left*) and JStor Daily (*highly factual*, *center-left*). This reflects both the dynamic nature of news media as well as the limits of the classification methodology used.

researchers from four different cultural backgrounds: Canada, Germany, the Netherlands and the United States. Each expert conducted an independent review and classified each of the 413 items generated in Phase 1 as either *fake news* or *real news*. All items with a ¾ expert consensus *and* matching with the correct answer key (i.e., the source veracity category)—a total of 289 items—were selected for the next phase.[83] A full list of the expert judgements and inter-rater agreement can be found in Supplement S1.

Phase 3: Quality Control. As a final quality control before continuing to the psychometrics study, the initial item selection committee in combination with an extra third expert—who had not been previously exposed to any of the items—made a final selection of items from Phase 2. Applying a ⅔ expert consensus as cutoff, we selected 100 items (44 fake news, 56 real news), thus creating a fairly balanced item pool for empirical probing that hosted five times as many as the final scale that we aimed to construct—in keeping with conservative guidelines (Boateng et al., 2018; Weiner et al., 2012). A full list of the item sets selected per expert and expert agreement can be found in Supplement S1.

Implementation

Participants. In line with widespread recommendations to assess at least 300 respondents during initial scale implementation (Boateng et al., 2018; Clark & Watson, 1995, 2019; Comrey & Lee, 1992; Guadagnoli & Velicer, 1988) we recruited a community sample of 452 US residents (for a comprehensive sample description see Table 7.2.2.1). The study was carried out on Prolific Academic (https://www.prolific.co/), an established crowd-working platform which provides competitive data quality (Palan & Schitter, 2018; Peer et al., 2017). Based on the exclusion criteria laid out in the preregistration, we removed incomplete cases, participants who took either an unreasonably short or long time to complete the study (less than 8 minutes or more than 2 hours), participants who failed an attention

[83] We used ¾ as a criterion instead of 100% consensus because as experts we may be biased ourselves, therefore we accepted items where only one expert did not agree on as well. If less than 120 items would remain then the Phase 1 item generation process would be restarted.

check, underage participants, and participants who did not live in the United States, retaining 409 cases for data analysis. Of these, 225 participants (i.e., 55.01%) participated in the follow-up data collection eight months later (T2).

Participants received a set remuneration of 1.67 GBP (equivalent to US$ 2.32) for participating in the T1 questionnaire and 1.10 GBP (equivalent to US$ 1.53) for T2.

Procedure, Measures, Transparency and Openness

The preregistrations for T1 and T2 are available on AsPredicted (https://aspredicted.org/blind.php?x=dw2kk9; https://aspredicted.org/blind.php?x=fh92dw; any deviations can be found in Supplement S2). Raw and clean datasets, as well as analysis scripts in R, can be found on the OSF repository (https://osf.io/r7phc/).

Participants took part in a preregistered online survey. After providing informed consent, participants had to categorize the 100 news headlines from Phase 3 (i.e., the items that were retained after the previous three phases) in two categories: *Fake/Deceptive* and *Real/Factual*.[84] Participants were told that each headline had only one correct answer (see preregistration for exact survey framing).

After completing the 100-item categorization task, participants completed the 18 items from the DEPICT inventory (Maertens et al., 2021), a 30-item COVID-19 fact-check test (Pennycook, McPhetres, et al., 2020), the Bullshit Receptivity scale (BSR; Pennycook et al., 2015), the Conspiracy Mentality Questionnaire (CMQ; Bruder et al., 2013), the Cognitive Reflection Test (CRT; Frederick, 2005), a COVID-19 compliance index (sample item: “I kept a distance of at least two meters to other people.”, 1—*does not apply at all*, 4—*applies very much*), and a demographics questionnaire (see Table 7.2.2.1 for an overview). Finally, participants were debriefed. Eight months later, the participants were recruited again for a test-retest follow-up survey.[85] In the follow-up survey, after participants provided informed

[84] All headlines can be found in Supplement S1.
[85] We chose to have a follow-up to be able to measure changes in the MIST score over the medium long-term. We found a period of eight months fitting for this purpose.

consent to participate, the final 20-item MIST was administered, the same COVID-19 fact-check test (Pennycook, McPhetres, et al., 2020) and CMQ (Bruder et al., 2013) were repeated, a new COVID-19 compliance index was administered, and finally a full debrief was presented. The complete surveys are available in the OSF repository: https://osf.io/r7phc/.

The full study received IRB approval from the Psychology Research Ethics Committee of the University of Cambridge (PRE.2019.108).

Analytical Strategy

To extract the final MIST-20 and MIST-8 scales from the pre-filtered MIST-100 item pool, we followed an item selection decision tree, which can be found in Supplement S3. Specifically—after ascertaining the general suitability of the data for such procedures—the following EFA- and IRT-based exclusion criteria were employed: (1) factor loadings below .40 (Clark & Watson, 2019; Ford et al., 1986; Hair et al., 2010; Rosellini & Brown, 2021); (2) cross-loadings above .30 (Boateng et al., 2018; Costello & Osborne, 2005); (3) communalities below .4 (Carpenter, 2018; Fabrigar et al., 1999; Worthington & Whittaker, 2006); (4) Cronbach's α reliability analysis; (5) differential item functioning (DIF) analysis (Holland & Wainer, 1993; Nguyen et al., 2014; Reise et al., 1993); (6) item information function (IIF) analysis. Finally, we sought to establish initial evidence for construct validity (Cronbach & Meehl, 1955). To do this, we investigated the associations between the MIST scales and the DEPICT deceptive headline recognition test (Maertens et al., 2021) as well as the COVID-19 fact-check (Pennycook, McPhetres, et al., 2020; concurrent validity). We further examined additional predictive accuracy of the MIST in accounting for variance in DEPICT and fact-check scores above and beyond the CMQ (Bruder et al., 2013), BSR (Pennycook et al., 2015) and CRT (Frederick, 2005; incremental validity).

7.3.2 Results

Item Selection

As a prerequisite for subsequent factor analyses, the data's factorability was tested via the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy and Bartlett's Test of Sphericity using R and the *EFAtools* package (Steiner & Grieder, 2020). Both tests indicated excellent data suitability (Bartlett's $\chi^2 = 12896.84$, $df = 4950$, $p < .001$; $KMO = .831$) according to established guidelines (Carpenter, 2018; Tabachnick & Fidell, 2007). Using parallel analysis with the *psych* package (Revelle, 2021) we aimed to select a parsimonious factor structure with each factor reflecting (i) an eigenvalue >1 and (ii) an eigenvalue larger than the simulated value (above the line of randomly generated data). Parallel analysis indicated five factors that matched our criteria, with two factors explaining most of the variance (eigenvalues: $F_1 = 10.89$, $F_2 = 7.82$, $F_3 = 1.89$, $F_4 = 1.42$, $F_5 = 1.23$). This fitted with our theoretical model of two factors (fake news detection and real news detection), with one potential higher-order news discernment factor, and two potential response bias factors. An exploratory factor analysis without rotation using the *EFAtools* package (Steiner & Grieder, 2020) indicated that both for the two-factor structure and the five-factor structure, the first two factors were specifically linked to the real news items and the fake news items respectively, while the other three factors did not show a pattern easy to interpret and in general low factor loadings (< .30). See Supplement S4 for a pattern matrix.

As we set out to create a measurement instrument for two distinct abilities, real news detection and fake news detection, we continued with a two-factor exploratory factor analysis, employing principal axis factoring and varimax rotation using the *psych* package (Revelle, 2021). Theoretically we would expect a balancing out of positive and negative correlations between the two factors: positive because of the underlying veracity discernment ability, and negative because of the response biases. We chose an orthogonal rotation instead

of an oblique rotation as we wanted to separate out fake news detection and real news detection as cleanly as possible.

Three iterations were needed to remove all items with a factor loading under .40 (43 items were removed). After this pruning, no items showed cross-loadings larger than .30. Communality analysis using the three-parameter logistic model function in the *mirt* package (Chalmers, 2012) with 50% guessing chance (c = .50) indicated two items with a communality lower than .40 after one iteration. These items were removed. No further iterations yielded any additional removals. A final list of the communalities can be found in Supplement S5. Cronbach's α reliability analysis with the *psych* package was used to remove all items that have negative effects ($\Delta\alpha > .001$) on the overall reliability of the test (Revelle, 2021). No items had to be removed based on this analysis. Differential item functioning using the *mirt* package was used to explore whether differences in gender or ideology would alter the functioning of the items (Chalmers, 2012). None of the items showed differential functioning for gender or ideology.

Finally, using the three-parameter logistic model IRT functions in the *mirt* package (Chalmers, 2012), we made a selection of the 20 best items (10 fake, 10 real) and the 8 best items (4 fake, 4 real), resulting in the MIST-20 and the MIST-8 respectively. These items were selected based on their discrimination and difficulty values, where we aimed to select a diverse set of items that have a high discrimination ($a \geq 2.00$ for the MIST-20, $a \geq 3.00$ for the MIST-8) but yet have a wide range of difficulties (b = [-0.50, 0.50], for each ability), while keeping the guessing parameter at 50% chance (c =.50). A list of the IRT coefficients and plots can be found in Supplement S1 and Supplement S6, respectively. See Figure 7.3.2.2 for a MIST-20 item trace line plot, and Figure 7.3.2.3 for a multidimensional plot of the MIST-20 IRT model predictions. The final items that make up the MIST-20 and MIST-8 are shown in Table 7.3.2.1. An overview of different candidate sets and how they performed can

be found in Supplement S7. The full analysis script can be found in the OSF repository: https://osf.io/r7phc/.

Table 7.3.2.1
Final Items Selected for MIST-20 and MIST-8

Item #	*a*	*b*	Content
Fake News			
MIST_14	**3.50**	**0.53**	**Government Officials Have Manipulated Stock Prices to Hide Scandals**
MIST_28	2.69	0.06	The Corporate Media Is Controlled by the Military-industrial Complex: The Major Oil Companies Own the Media and Control Their Agenda
MIST_20	**3.26**	**-0.20**	**New Study: Left-Wingers Are More Likely to Lie to Get a Higher Salary**
MIST_34	3.42	-0.25	The Government Is Manipulating the Public's Perception of Genetic Engineering in Order to Make People More Accepting of Such Techniques
MIST_15	2.34	-0.40	Left-Wing Extremism Causes 'More Damage' to World Than Terrorism, Says UN Report
MIST_7	**2.57**	**-0.45**	**Certain Vaccines Are Loaded with Dangerous Chemicals and Toxins**
MIST_19	2.00	-0.55	New Study: Clear Relationship Between Eye Color and Intelligence
MIST_33	**5.60**	**-0.76**	**The Government Is Knowingly Spreading Disease Through the Airwaves and Food Supply**
MIST_10	2.64	-1.02	Ebola Virus 'Caused by US Nuclear Weapons Testing', New Study Says
MIST_13	2.86	-1.30	Government Officials Have Illegally Manipulated the Weather to Cause Devastating Storms
Real News			
MIST_50	**3.12**	**0.38**	**Attitudes Toward EU Are Largely Positive, Both Within Europe and Outside It**
MIST_82	2.22	0.31	One-in-Three Worldwide Lack Confidence in NGOs
MIST_87	2.25	0.14	Reflecting a Demographic Shift, 109 US Counties Have Become Majority Nonwhite Since 2000
MIST_65	2.36	-0.03	International Relations Experts and US Public Agree: America Is Less Respected Globally
MIST_60	**3.39**	**-0.09**	**Hyatt Will Remove Small Bottles from Hotel Bathrooms by 2021**
MIST_73	2.43	-0.14	Morocco's King Appoints Committee Chief to Fight Poverty and Inequality
MIST_88	**2.79**	**-0.31**	**Republicans Divided in Views of Trump's Conduct, Democrats Are Broadly Critical**
MIST_53	2.12	-0.37	Democrats More Supportive than Republicans of Federal Spending for Scientific Research
MIST_58	**8.59**	**-0.60**	**Global Warming Age Gap: Younger Americans Most Worried**
MIST_99	2.26	-0.83	US Support for Legal Marijuana Steady in Past Year

Note. Items in bold are items included in the short version of the test (MIST-8). *a* = discrimination parameter. *b* = difficulty parameter.

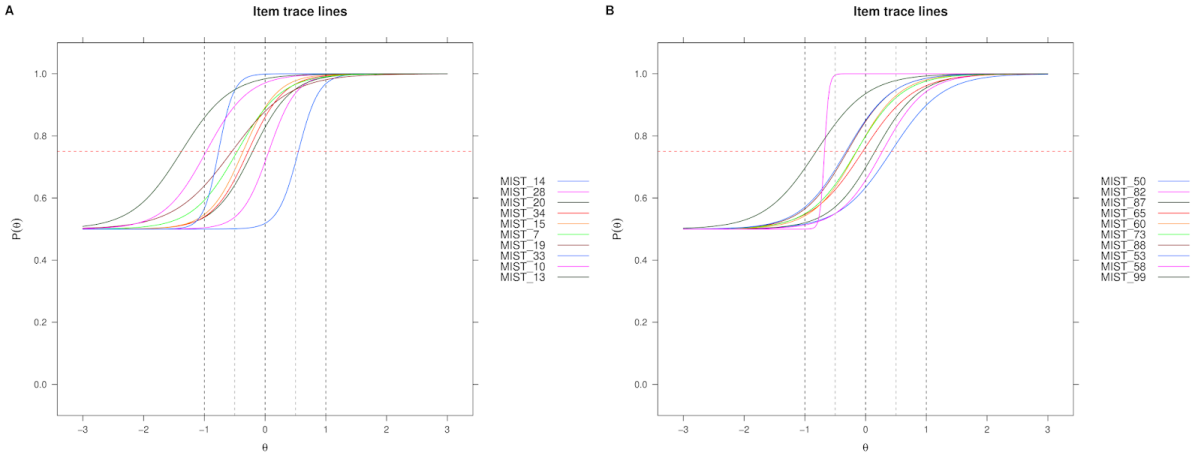

Figure 7.3.2.2. Item trace lines for MIST-20 items, for the fake news items in Panel A and real news items in Panel B. The horizontal axis represents the latent skill (θ). The vertical axis represents the probability of correctly responding to the item.

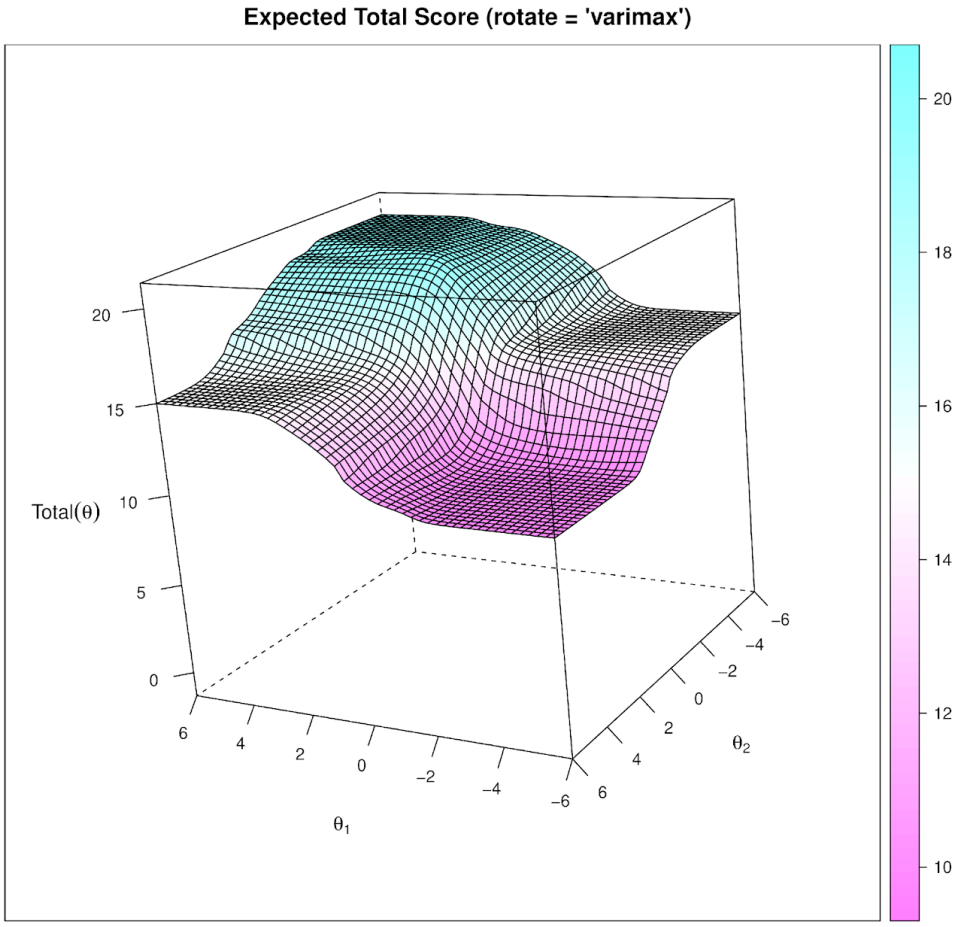

Figure 7.3.2.3. Multi-dimensional IRT plot representing the final MIST-20 test. The vertical axis represents the discernment skill, the other two axes represent the real news detection (θ_1) and fake news detection (θ_2).

Reliability

Inter-item correlations show a good internal consistency both for the MIST-8 (IIC_{min} = .20, IIC_{max} = .27) and for the MIST-20 (IIC_{min} = .22, IIC_{max} = .29). Item-total correlations also show a good reliability both for the MIST-8 (ITC_{min} = .44, ITC_{max} = .53) and for the MIST-20 (ITC_{min} = .31, ITC_{max} = .54).

Looking further into the MIST-20, we analyze the reliability of veracity discernment (**V**; M = 15.71, SD = 3.35), real news detection (**r**; M = 7.62, SD = 2.43), and fake news detection (**f**; M = 8.09, SD = 2.10). In line with the guidelines by Revelle and Condon (2019), we calculate a two-factor McDonald's ω (McDonald, 1999) as a measure of internal consistency using the *psych* package (Revelle, 2021), and find a good reliability for the general scale and the two facet scales (ω_g = 0.79, ω_{F1} = 0.78, ω_{F2} = 0.75). Also using the *psych* package (Revelle, 2021), we calculated the variance decomposition metrics as a measure of stability, finding that F1 explains 14% of the total variance and F2 explains 12% of the total variance. Relatively, of all variance explained, 53% comes from F1 (**r)** and 47% comes from F2 (**f**), demonstrating a good balance between the two factors.

Finally, test-rest reliability analysis indicates that MIST scores are moderately positively correlated over a period of eight-to-nine months ($r_{T1,T2}$ = 0.58).[86]

Validity

To assess initial validity, we examined the associations between the MIST scales and two scales that have been used regularly in previous misinformation research: the COVID-19 fact-check by Pennycook, McPhetres, et al. (2020) as well as the DEPICT test by Maertens et al. (2021), expecting high correlations (r > .50; concurrent validity) and *additional* variance explained as compared to the existing CMQ, BSR, and CRT scales (incremental validity;

[86] It must be noted that at T2 participants only completed the 20-item MIST, while at T1 participants had to categorize 100 items, with a slightly different question and response framings (see full Qualtrics lay-outs and question framings in the OSF repository: https://osf.io/r7phc/). We expect the actual test-retest correlation to be higher.

Clark & Watson, 2019; Meehl, 1978). As can be seen in Table 7.3.2.4, we found that the MIST-8 displays a medium-to-high correlation with the fact-check ($r_{\text{fact-check,MIST-20}} = .49$) and DEPICT test ($r_{\text{DEPICT,MIST-20}} = .45$), while the MIST-20 shows a large positive correlation with both the fact-check ($r_{\text{fact-check,MIST-20}} = .58$) and the DEPICT test ($r_{\text{DEPICT,MIST-20}} = .50$). Using a linear model, we found that the explained variance in the fact-check indicates that the MIST-20 can explain 33% (adjusted R^2) of variance by itself. The CMQ, BSR, and CRT combined account for 19%. Adding the MIST-20 on top provides an incremental 18% of explained variance (adjusted $R^2 = 0.37$). The MIST-20 is the strongest predictor in the combined model ($t(404) = 10.82$, $p < .001$, $\beta = 0.49$, 95% CI [0.40, 0.57]). For the DEPICT test we found that the CMQ, BSR, and CRT combined explain 12% of variance in deceptive headline recognition and 26% when the MIST-20 is added ($\Delta R^2 = 0.14$), while the MIST-20 alone explains 25%. For the DEPICT test we found the MIST-20 to be the only significant predictor in the combined model ($t(404) = 8.94$, $p < .001$, $\beta = 0.43$, 95% CI [0.34, 0.53]).[87]

Table 7.3.2.4

Incremental Validity of MIST-8 and MIST-20 With Existing Measures

CV19 Fact-Check ~			
	r	adjusted R^2	ΔR^2
MIST-8	.49	.24	
MIST-20	.58	.33	
CMQ + BSR + CRT		.19	
CMQ + BSR + CRT + MIST-8		.30	.11***
CMQ + BSR + CRT		.19	
CMQ + BSR + CRT + MIST-20		.37	.18***
DEPICT ~			
MIST-8	.45	.20	
MIST-20	.50	.25	
CMQ + BSR + CRT		.12	
CMQ + BSR + CRT + MIST-8		.22	.11***
CMQ + BSR + CRT		.12	
CMQ + BSR + CRT + MIST-20		.26	.14***

* $p < .05$, ** $p < .01$, *** $p < .001$

[87] Full model output for the MIST-8 and MIST-20 linear models can be found in Supplement S8. Full analysis scripts can be found in the OSF repository: https://osf.io/r7phc/.

7.3.3 Discussion

In Study 7A, we generated 413 news items using GPT-2 automated item generation for fake news, and trusted sources for real news. Through two independent expert committees, we reduced the item pool to 100 items (44 fake and 56 real) and applied multidimensional item-response theory in combination with factor analysis to reduce the item set to the 20 best items for the MIST-20 and the 8 best items for the MIST-8. We found that the final items demonstrate good reliability. In an initial test of validity, we found strong concurrent validity for both the MIST-8 and the MIST-20 as evidenced through their strong associations with the COVID-19 fact-check and the DEPICT deceptive headline recognition test. Moreover, we found that both the MIST-20 and the MIST-8 outperformed the combined model of the CMQ, BSR, and CRT, when explaining variance in fact-check and DEPICT scores, evidencing incremental validity. This study provides the first indication that both the MIST-20 and MIST-8 are psychometrically sound, and can explain and test misinformation susceptibility above and beyond the existing scales.

Study 7B

7.4 Validation—Confirmatory Analyses, Nomological Net, and National Norms

Study 7B sought to consolidate and extensively test the psychometric soundness of the newly developed MIST-20. Across four large nationally representative samples from two countries (US, UK) and three different recruitment platforms (CloudResearch, Prolific, and Respondi) we pursued three goals. First, we used structural equation modelling and reliability analyses to probe the structural stability, model fit, and internal consistency of the MIST across different empirical settings. Second, we built an extensive nomological network and examined correlation patterns as well as predictive power of the MIST to demonstrate convergent, discriminant, and incremental validity. Third, we capitalized on the representativeness of our samples to derive national norms for the general population (UK, US) and specific demographic (UK, US) and geographic subgroups (US).

7.4.1 Methods

Participants

As part of our validation study, we collected data from four nationally representative samples (N_{total} = 8,310, N_{clean} = 6,461).[88] Sample 2A was a US representative sample (N = 3,692) with interlocking age and gender quota (i.e., each category contains a representative relative proportion of the other category) accessed through *Respondi,* an ISO-certified international organization for market and social science research (for previous applications see e.g., Dür & Schlipphak, 2021; Heinsohn et al., 2019; Roozenbeek, Freeman, et al., 2021). After excluding incomplete cases and participants outside of the quota, 3,479 participants were considered for analysis. Sample 2B was a US representative sample with age, ethnicity, and gender quota (N = 856) recruited through *CloudResearch* (formerly *TurkPrime*), an

[88] Surveys 2A, 2C, and 2D were designed as part of a separate research project which featured the MIST-20 as an add-on. Survey 2B was designed specifically for this project.

online research platform similar to MTurk but with additional validity checks and more intense participant pool controls (Buhrmester et al., 2018; Litman et al., 2017). After excluding, as preregistered, all participants who failed an attention check, were underage, did not reside in the United States, did not complete the entire study, or were a second-time participant, 510 participants remained.[89] Sample 2C was a UK representative sample (N = 2,517) based on interlocking age and gender quota recruited through *Respondi*. After excluding incomplete cases and participants outside of our quota criteria, 1,227 participants were retained. Lastly, sample 2D was a UK representative sample (N = 1,396) with age and gender quota recruited through *Prolific.* Excluding all entries that fell outside of our quota criteria and all incomplete entries resulted in an analysis sample of 1,245 participants.

In line with the best practices for scale development to recruit at least 300 participants per sample (Boateng et al., 2018; Clark & Watson, 1995, 2019; Comrey & Lee, 1992; Guadagnoli & Velicer, 1988), as well as being highly powered (power = .90, α = .05) to detect the smallest effect size of interest (r = .10, needed N = 1,046; Anvari & Lakens, 2021; Funder & Ozer, 2019; Götz, Gosling, et al., 2021), Sample 2A, 2C, and 2D exceed the size requirements. Sample 2B was highly powered (power = .90, α = .05) to detect effect sizes r of .15 (needed N = 463). Power analyses were completed using the *pwr* package in R (Champely et al., 2021).

Detailed demographic breakdowns of all samples are shown in Table 7.2.2.1.

Procedure and Measures

All participants were invited to take part in an online survey through the respective research platforms. After providing informed consent, all participants provided basic demographics information, completed the MIST-20 and—depending on their sample group—a select set of additional psychological measures (for a detailed description of all

[89] This is a slight deviation from the preregistration, where we did not mention incomplete entries or second entries. We also removed participants who failed *any* attention check instead of *both*. Both were done to ensure data quality.

constructs assessed in each sample group, see Table 7.2.2.1). All participants received financial compensation in accordance with platform-specific remuneration standards and guidelines on ethical payment at the University of Cambridge. Participants in Sample 2A, 2B, and 2C were paid by the sampling platform directly, while participants in Sample 2D received 2.79 GBP for a 25-minute survey (6.70 GBP per hour). All data collections were approved by the Psychology Research Ethics Committee of the University of Cambridge (PRE.2019.108, PRE.2020.034, PRE.2020.086, PRE.2020.120).

Analytical Strategy

We adopted a three-pronged analytical strategy. First, we computed reliability estimates and conducted confirmatory factor analyses for each sub-sample, seeking to reproduce, consolidate, and evaluate the higher-order model derived in Study 7A. Second, in an effort to establish construct validity (Cronbach & Meehl, 1955; Strauss & Smith, 2009) we pooled the constructs assessed across our four validation samples to build a comprehensive, theory-driven and preregistered (Sample 2B) nomological network. As such, we cast a wide net and included 1) concepts that should be meaningfully positively correlated with MIST scores (convergent validity; i.e., *DEPICT Balanced Short Form*; Maertens et al., 2021; *Go Viral! Balanced Item Set*; Basol et al., 2021) expecting a high positive Pearson *r* correlation ([0.50, 0.80]); 2) concepts that should be clearly distinct from the MIST (discriminant validity; i.e., *Bullshit Receptivity Scale*; BSR; Pennycook et al., 2015; *Conspiracy Mentality Questionnaire*; CMQ; Bruder et al., 2013) expecting a low-to-medium negative correlation with the MIST (Pearson $r = [-0.50, -0.20]$) and 3) an array of prominent psychological constructs of general interest (i.e., personality traits, attitudes and cognitions including the *Big Five*, *Dark Tetrad*, *Moral Foundations*, *Social Dominance Orientation*, *Ecological Dominance Orientation*, religiosity, self-esteem, political cynicism, numeracy and trust in various public institutions and social agents). Third, we leveraged the size and

representativeness of our samples to establish norm tables for the US and UK general populations as well as specific demographic and geographical subgroups.

7.4.2 Results

Model Fit

For each sample, we employed structural equation modelling to assess model fit—examining both a basic first-order model with two distinct factors (i.e., real news detection, fake news detection) as well as a theoretically-derived higher-order model (Markon, 2019; Thurstone, 1944) in which both first-order factors load onto a general second-order veracity discernment factor. We then calculated reliability estimates using internal consistency measures (inter-item correlations, item-total correlations, and MacDonald's omega). We used the *lavaan* package for structural equation modelling in R (Rosseel, 2012).

In keeping with our theoretical conceptualization of the MIST—with a general ability factor of veracity discernment, as well as two superordinate factors capturing real news and fake news detection, respectively, we fitted a higher-order model (Markon, 2019; Thurstone, 1944) in which both first-order factors load onto a general second-order veracity discernment factor (see Figure 7.4.2.1). We first did this with Sample 2A (representative US sample from *Respondi*). Consistent with conventional guidelines (RMSEA/SRMR < .10 = acceptable; < .06 = excellent; CFI/TLI > .90 = acceptable; > .95 = excellent; Clark & Watson, 2019; Finch & West, 1997; Hu & Bentler, 1999; Pituch & Stevens, 2015; Schumacker et al., 2015), the model fits the data well (MIST-20: CFI = .90, TLI = .89, RMSEA = .041, SRMR = .040; MIST-8: CFI = .97, TLI = .95, RMSEA = .030, SRMR = .025). We note that the χ^2 goodness of fit test was significant—signaling lack of fit (MIST-20: $\chi^2 = 1021.86$, $p < .001$; MIST-8: ; $\chi^2 = 72.74$, $p < .001$). However, this should be interpreted with caution, as the χ^2 is a test of perfect fit and very sensitive to sample size. As such, as sample sizes approach 500, χ^2 is

usually significant even if the differences between the observed and model-implied covariance matrices are trivial (Bentler & Bonett, 1980; Curran et al., 2003; Rosellini & Brown, 2021). Taken together, the findings thus suggest good model fit for the theoretically derived higher-order model.

Importantly this model also yielded better fit than a traditional basic first-order model (with two distinct fake news and real news factors; MIST-20: $\chi^2 = 1027.17$, $p < .001$, *CFI* = 0.90, *TLI* = 0.89, *RMSEA* = 0.041, *SRMR* = 0.041; MIST-8: $\chi^2 = 99.46$, $p < .001$, *CFI* = 0.95, *TLI* = 0.93, *RMSEA* = 0.035, *SRMR* = 0.035). A likelihood-ratio test of the higher-order model versus the first-order model was significant for both the MIST-20 and for the MIST-8 (MIST-20: $\Delta\chi^2 = 5.35$, $p = .021$, MIST-8: $\Delta\chi^2 = 26.29$, $p < .001$), indicating a better fit for the higher-order model.

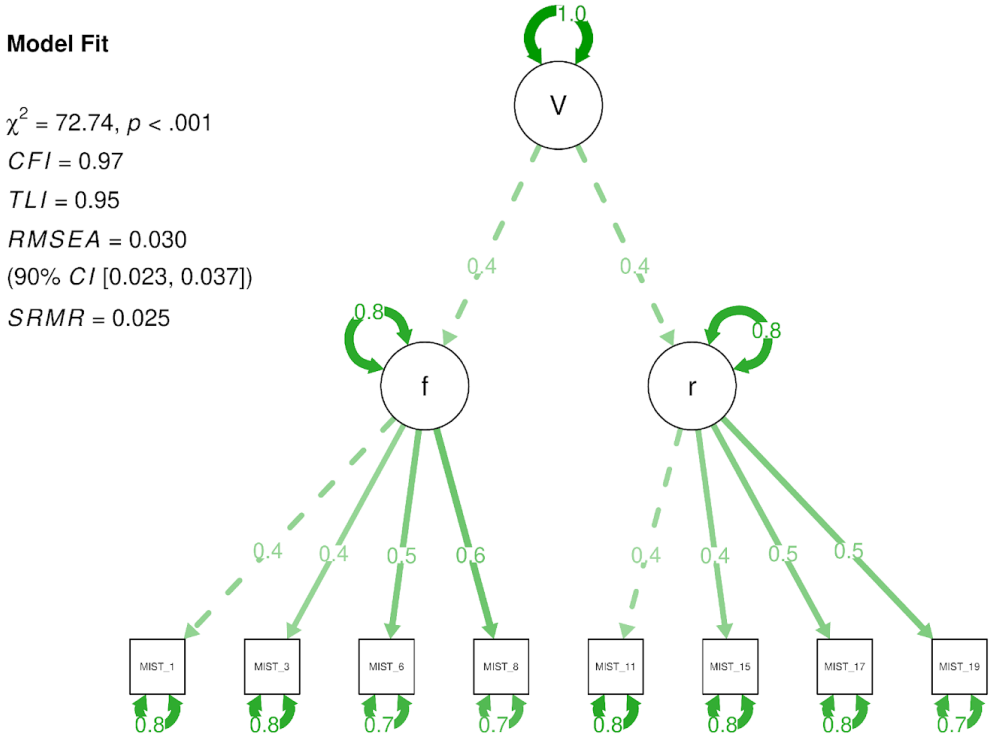

Figure 7.4.2.1. Plot of higher order MIST-8 SEM model in Sample 2A ($N = 3{,}479$). *V* represents veracity discernment, *f* fake news detection, and *r* real news detection.

Sample comparison. Across all four samples, we successfully reproduced the original higher-order model, with parameters indicating good fit, as well as good internal consistency in all four samples (see Table 7.4.2.2 for a complete overview).[90] A similar fit is found between the US *Respondi* and UK *Respondi* samples, indicating that the MIST works similarly in the UK as it does in the US. Meanwhile, larger differences are found between the US *Respondi* and the US *CloudResearch samples*, and between the UK *Respondi* and the UK *Prolific* samples, indicating that sampling platform plays a larger role than nationality when administering the MIST even when using representative quota sampling.

Table 7.4.2.2
Model Fit Overview

MIST-20												
Samp.	Plat.	Pop.	χ^2	*p*	*CFI*	*TLI*	*RMSEA*	95% CI		*SRMR*	ω_{tot}	3F
								LL	*UL*			
2A	R	US	1021.86	< .001	0.90	0.89	0.041	0.039	0.044	0.040	0.76	*
2B	C	US	264.66	< .001	0.92	0.91	0.035	0.027	0.043	0.051	0.75	.
2C	R	UK	473.56	< .001	0.91	0.90	0.041	0.037	0.046	0.049	0.81	***
2D	P	UK	432.12	< .001	0.86	0.85	0.038	0.034	0.042	0.045	0.70	***

MIST-8												
Samp.	Plat.	Pop.	χ^2	*p*	*CFI*	*TLI*	*RMSEA*	95% CI		*SRMR*	ω_{tot}	3F
								LL	*UL*			
2A	R	US	72.74	< .001	0.97	0.95	0.030	0.023	0.037	0.025	0.57	***
2B	C	US	30.32	.048	0.96	0.94	0.036	0.003	0.058	0.040	0.58	*
2C	R	UK	64.13	< .001	0.94	0.91	0.045	0.033	0.058	0.040	0.62	***
2D	P	UK	46.91	< .001	0.93	0.90	0.037	0.023	0.050	0.035	0.55	***

Note. Total N = 6,461. Samp = sample. Plat = sampling platform. Pop = sample population. CI = confidence interval; LL = lower limit; UL = upper limit. 3F reflects whether the three-factor (higher-order) model provided better fit than the two-factor (two-order) model; . = descriptively better fit but not significant; * $p < .05$, ** $p < .01$, *** $p < .001$.

Nomological Network[91]

Convergent Validity. As preregistered, in Sample 2B—which was the sample we primarily relied on in constructing the nomological network—the correlation between the general MIST-20 score and the DEPICT Balanced Short Form measure (Maertens et al.,

[90] Supplement S9 includes model plots for both the MIST-20 and MIST-8 for all samples.
[91] This section focuses on the nomological network of the general ability factor (veracity discernment) of the MIST-20. However, we have also constructed nomological networks for the subcomponents of the MIST as well as the MIST-8. For parsimony's sake, these are reported in Supplements S10-S12.

2021) was found to be positive and medium-to-large, with a significant Pearson correlation of .54 (95% CI [.48, .60], p < .001).[92] The MIST-20 correlation with the *Go Viral!* inventory (Basol et al., 2021) was under the estimated value but was significantly correlated with a Pearson correlation of .26 (95% CI [.18, .34], $p < .001$). Similarly, regarding incremental validity, the additional explained variance in the DEPICT Balanced Short Form measure above and beyond the CMQ and the BSR is at the upper side of our prediction, with an additional 20% of variance explained, whereas with 3% it is under the predicted value for the *Go Viral!* inventory.[93] For a more detailed account see Supplement S13. In addition, in Sample 2A, we measured belief in COVID-19 myths, which was significantly positively correlated and within the preregistered strength of convergent validity measures ($r = -.51$, 95% CI [-.55, -.47], $p < .001$). See Figure 7.4.2.3 for a regression plot.

Discriminant Validity. As preregistered for Sample 2B, the MIST-20 was moderately negatively correlated with the BSR ($r = -.21$, [-.29, -.13], $p < .001$) and the CMQ ($r = -.38$ [-.45, -.30], $p < .001$). Overall, the correlational pattern of our nomological network supports the construct validity of the MIST, with the MIST being more strongly correlated with the convergent measures than with the discriminant measures (Campbell & Fiske, 1959; Rosellini & Brown, 2021).

CRT (Sample 2A). In line with other studies finding a role for the CRT in misinformation detection (e.g., Pennycook & Rand, 2019), we found a significant correlation between the MIST score and the cognitive reflection test, or CRT ($r = .29$, 95% CI [.26, .32], $p < .001$).

AOT (Sample 2A). We found an even larger significant correlation between the MIST score and actively open-minded thinking or AOT ($r = .49$, 95% CI [.46, .51], $p < .001$).

[92] See https://aspredicted.org/blind.php?x=4v9eq7 for the preregistration (Sample 2B).
[93] It has to be noted that the Go Viral inventory is not a validated measurement instrument. Results should be interpreted in light of this.

BFI (Sample 2B). Contrary to our preregistered exploratory hypotheses, in Sample 2B the MIST-20 score was not significantly correlated with openness, $r = .02$, 95% CI [-.06, -.11], $p = .594$, and agreeableness was *not* negatively correlated with distrust **d**, $r = .05$, 95% CI [-.04, .14], $p = .255$. The MIST-20 score was also not significantly correlated with agreeableness ($r = .05$, 95% CI [-.04, .14], $p = .271$) or extraversion ($r = -.07$, 95% CI [-.15, .02], $p = .141$), but did significantly correlate with conscientiousness ($r = .10$, 95% CI [.02, .19], $p = .020$) and neuroticism ($r = -.14$, 95% CI [-.23, -.06], $p = .001$).

DT (Sample 2B). The MIST-20 score was negatively correlated with each of the four Dark Tetrad traits: *Machiavellianism* ($r = -.09$, 95% CI [-.17, -.00], $p = .047$), *narcissism* ($r = -.26$, 95% CI [-.34, -.18], $p < .001$), *psychopathy* ($r = -.30$, 95% CI [-.37, -.22], $p < .001$), and *sadism* (-.22, 95% CI [-.30, -.12], $p < .001$). However, contrary to our preregistered exploratory hypothesis, machiavellianism was *not* negatively correlated with naïvité **n**, $r = .16$, 95% CI [.07, .24], $p < .001$.

Trust Measures (Sample 2B). In line with our preregistered exploratory hypotheses, we found that the MIST score *was* correlated with trust in science, $r = .33$, 95% CI [.25, .41], $p < .001$, scientists, $r = .36$, 95% CI [.28, .43], $p < .001$, and mainstream media, $r = .18$, 95% CI [.09, .26], $p < .001$. In addition, we found that trust in doctors, $r = .36$, 95% CI [.28, .43], $p < .001$, journalists, $r = .19$, 95% CI [.11, .27], $p < .001$, and officials, $r = .09$, 95% CI [.00, .17], $p = .049$, were significantly positively correlated, while trust in the government, $r = -.11$, 95% CI [-.20, -.02], $p = .012$, was significantly negatively correlated with the MIST-20. We found no significant correlation for either of the two trust-in-politicians scales, $r_a = -.06$, 95% CI [-.14, .03], $p = .210$, $r_b = .07$, 95% CI [-.02, .15], $p = .131$.

Additional Associations. For a summary and discussion of the exploratory analyses of MFQ, SDO, EDO, religiosity, and demographics, please see Supplement S14.

Detailed summary figures separated by outcome category are available in Supplements S10-S12.

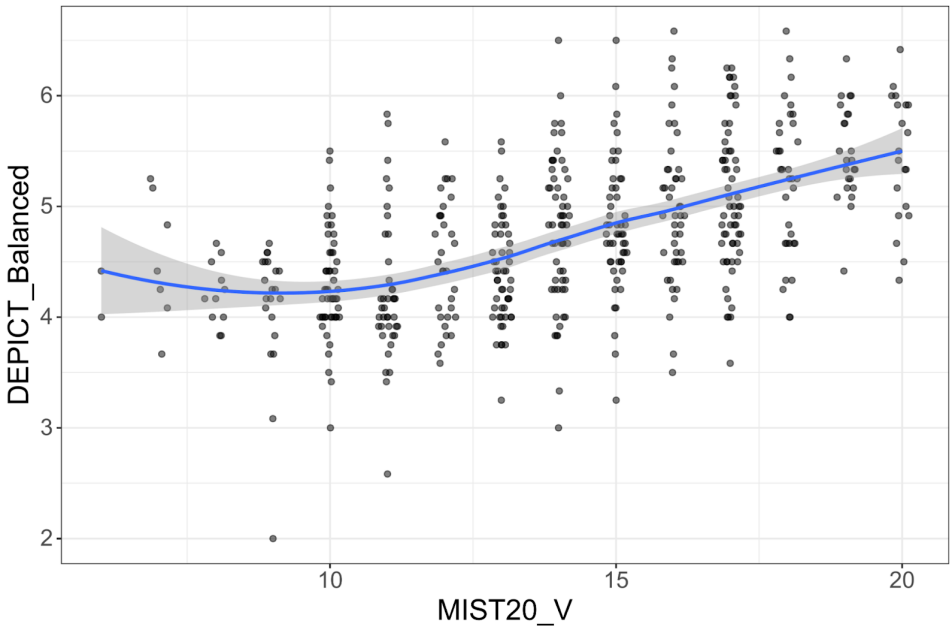

Figure 7.4.2.3. Regression of DEPICT Balanced Short Form score on MIST-20 veracity discernment score, with 95% confidence band, in Sample 2B ($N = 510$).

National Norms

We used the *Respondi* samples for each country (i.e., Sample 2A for the US and Sample 2C for the UK) to generate norm tables for general veracity discernment as well as fake news and real news detection.[94] As can be gleaned from Table 7.4.2.4, the norms for both countries were very similar with minor deviations of single score points, further corroborating evidence for the cross-cultural validity of the MIST. Table 7.4.2.5 exhibits norms for the general US population.

Full norm tables for the US and the UK, including specific norms based on age (US, UK) and geography (US; i.e., 9 census divisions, 4 census regions), as well as means and

[94] We chose to create the norm tables based on the Respondi samples instead of pooling all samples as through recent projects we found some evidence indicating that Respondi samples provide more representative levels of *numeracy*, *education*, and *ideology* than Prolific, and our experience with CloudResearch is limited.

standard deviations per item, including a per-item comparison between Democrats (US)/liberals (UK) and Republicans (US)/conservatives (UK), are available in Supplement S15.

Table 7.4.2.4

MIST Norm Score Comparison Between US and UK Sample

Scale	Sample	*Minimum*	1st *Quartile*	*Median*	*Mean*	3rd *Quartile*	*Maximum*
MIST-8							
	US	0	4	6	6	7	8
	UK	0	4	5	5	7	8
MIST-20							
	US	4	11	14	14	17	20
	UK	4	11	13	13	16	20

Table 7.4.2.5

MIST-20 General Population Norms for the United States (N = 3,479)

V (Veracity Discernment)		f (Fake News Detection)		r (Real News Detection)	
Percentile	Score	Percentile	Score	Percentile	Score
0%	4	0%	0	0%	0
5%	8	5%	3	5%	2
10%	9	10%	4	10%	3
15%	10	15%	5	15%	4
20%	10	20%	5	20%	4
25%	11	25%	6	25%	5
30%	12	30%	7	30%	5
35%	12	35%	7	35%	6
40%	13	40%	7	40%	6
45%	14	45%	8	45%	7
50%	14	50%	8	50%	7
55%	15	55%	8	55%	7
60%	15	60%	9	60%	7
65%	16	65%	9	65%	8
70%	16	70%	9	70%	8
75%	17	75%	9	75%	8
80%	17	80%	10	80%	9
85%	18	85%	10	85%	9
90%	19	90%	10	90%	10
95%	19	95%	10	95%	10
100%	20	100%	10	100%	10

7.4.3 Discussion

In Study 7B, we consolidated and expanded the psychometric properties of the MIST. First, we conducted confirmatory factor analyses across four nationally representative

samples from the US and the UK, consistently replicating the higher-order structure yielding good model fit and internal consistency for both the MIST-8 and the MIST-20. Next, we constructed an extensive nomological network of the MIST to assess construct validity (Cronbach & Meehl, 1955). As preregistered and similar to Study 7A, in Sample 2B we found a high correlation between the MIST score and the DEPICT misinformation inventory, supporting convergent validity. Similarly, in Sample 2A we found a medium-to-high negative correlation between the MIST-20 and a COVID-19 misinformation beliefs inventory further attesting to the measure's convergent validity. In addition, we demonstrated that both the MIST-8 and the MIST-20 explain considerable extra variance above the existing CMQ and BSR scales (MIST-20: $\Delta R^2 = 20\%$, MIST-8: $\Delta R^2 = 14\%$), indicating substantial incremental validity (Clark & Watson, 2019). Surprisingly, however, the correlations of each of the MIST, CMQ, and the BSR with the *Go Viral!* items were all low ($r < .30$). Nevertheless, the MIST-20 remained the single best predictor for the *Go Viral!* items, significantly improving the variance explained in a combined model on top of the CMQ and BSR measures ($\Delta R^2 = .03$). In terms of discriminant validity, as preregistered, in Sample 2B we observed moderate negative associations between the MIST-20 and the BSR as well as the CMQ.

All in all, the nomological network largely affirmed the preregistered relationship patterns—thus corroborating the MIST's construct validity—while at the same time demonstrating new insights that can be gained by using the MIST-20 measure, which may stimulate further research. Finally, we leveraged the large size and national representativeness of our validation samples to produce norm tables for the UK and US general populations as well as distinct demographic subgroups in the UK and the US and geographical subgroups in the US.

Study 7C

7.5 Application—A Nuanced Effectiveness Evaluation of a Popular Media Literacy Intervention

In Study 7C, we demonstrate how the MIST can be used in conjunction with the ***Verification done*** framework and norm tables.[95] We employ the MIST-8 in a simple within-groups pre-test post-test design with the *Bad News Game*, a major media literacy intervention played by over a million people (Roozenbeek & van der Linden, 2019a). The *Bad News Game* is based on inoculation theory (van der Linden & Roozenbeek, 2020), and its theoretical mechanisms as well as its effects have been replicated multiple times (see e.g., Maertens et al., 2021; Roozenbeek, Maertens, et al., 2021), making it a well-established intervention in the literature as a tool to reduce misinformation susceptibility. We therefore hypothesized that the intervention improves ***v****eracity discernment* (ability to accurately distinguish real news from fake news), ***r****eal news detection* (ability to correctly flag real news), and ***f****ake news detection* (ability to correctly tag fake news). In addition, we hypothesized that the *Bad News Game* decreases both **d**istrust (negative judgement bias or being hyper skeptical) and **n**aïvité (positive judgement bias or believing everything). We used norm tables to establish where the baseline MIST scores of our convenience sample lies.

[95] A MIST implementation guide explaining how researchers and practitioners can set up the MIST in their studies as well as how to calculate the *Verification done* (Vrf dn) scores can be found in Supplement S17. An example Qualtrics survey as well as a score calculation R script are available in the OSF repository: https://osf.io/r7phc/.

7.5.1 Methods

Participants

We collected data from an online community sample of 4,024 participants who played the *Bad News Game* (www.getbadnews.com) between 7th May 2020 and 29th July 2020 and agreed to participate in the in-game survey. After filtering out participants who did not complete the full study, had prior experience with the game, were underage, or entered the study multiple times, and lived outside of the United States, 421 participants remained.[96] Based on earlier studies evaluating the Bad News Game (Maertens et al., 2021; Roozenbeek, Maertens, et al., 2021), we aimed to be highly powered (power = .90, $\alpha = .05$) to detect a Cohen's *d* effect size of 0.250, which required a sample size of 338, which we exceed in this sample. The power was calculated using the R *pwr* package (Champely et al., 2021).

On average, participants were young (55.58% between 18–29, 32.30% between 30–49, 12.11% over 50), 52.02% identified as female (41.09% male, 6.89% other), and 86% had either a higher education degree or some college experience (see Table 7.2.2.1 for a complete demographics overview). The median ideology on a scale from 1 (*liberal*) to 7 (*conservative*) was 3 ($M = 2.88$, $SD = 1.39$), indicating a slightly left-leaning audience.

Procedure and Measures

Individuals who played the Bad News Game (Roozenbeek & van der Linden, 2019a) were invited to participate in the study. The Bad News Game (www.getbadnews.com) is a free online browser game in which players learn about six common misinformation techniques over the course of 15 minutes in a simulated social media environment (see Roozenbeek & van der Linden, 2019a, for a detailed discussion). In the current study—after providing informed consent—individuals completed the MIST-8 both before and after playing the Bad News Game. Participation was completely voluntary and no rewards,

[96] We restricted our sample to US residents, as we did not have a UK filter and have not yet validated the MIST in any other country.

monetary or otherwise, were offered. This study was approved by the Psychology Research Ethics Committee of the University of Cambridge (PRE.2020.120, PRE.2020.136).

Analytical Strategy

After contextualizing our findings by juxtaposing the sample's baseline findings to the US general national norms derived in Study 7B, we conducted repeated-measures *t*-tests for veracity discernment ($M = 6.23$, $SD = 1.53$) as well as the four subcomponents of the MIST—fake news detection ($M = 3.19$, $SD = 0.92$), real news detection ($M = 3.04$, $SD = 0.95$), distrust ($M = 0.31$, $SD = 0.63$), and naïvité ($M = 0.46$, $SD = 0.69$).

7.5.2 Results

Baseline

We found that our US convenience sample scored higher on the MIST than the US population average for **v**eracity discernment (see Study 7B; 1st $Quartile_{Population} = 4$, 1st $Quartile_{Sample} = 6$).[97]

Hypothesis Tests

V—Veracity Discernment. Contrary to our expectations, we did not find a significant effect of veracity discernment post-intervention as compared to pre-intervention ($M_{diff} = 0.11$, 95% CI [-0.09, 0.31], $t(839) = 1.06$, $p = .291$, $d = 0.088$, 95% CI [-0.103, 0.279]). See Figure 7.5.2.1, Panel A for a bar plot.

r—Real News Detection. While we found an effect of the intervention on real news detection, the effect was in the opposite direction of our prediction ($M_{diff} = -0.17$, 95% CI [-0.30, -0.03], $t(820) = -2.44$, $p = .015$, $d = -0.181$, 95% CI [-0.373, 0.010]). See Figure 7.5.2.1, Panel B for a bar plot.

[97] We found similar results when looking at Fake News Detection (1st $Quartile_{Population} = 2$, 1st $Quartile_{Sample} = 3$) and Real News Detection (1st $Quartile_{Population} = 2$, 1st $Quartile_{Sample} = 3$).

f—Fake News Detection. In line with our expectations, we did find a positive effect of the intervention on fake news detection (M_{diff} = 0.28, 95% CI [0.15, 0.40], $t(839) = 4.32$, $p < .001$, $d = 0.332$, 95% CI [0.139, 0.524]). See Figure 7.5.2.1, Panel C for a bar plot.

d—Distrust. Contrary to our hypothesis, we observed an increase in distrust (M_{diff} = 0.31, 95% CI [0.20, 0.42], $t(712) = 5.42$, $p < .001$, $d = 0.338$, 95% CI [0.146, 0.531]). See Figure 7.5.2.1, Panel D for a bar plot.

n—Naïvité. As hypothesized, we did find a significant reduction in naïvité after intervention (M_{diff} = -0.14, 95% CI [-0.23, 0.04], $t(840) = -2.90$, $p = .004$, $d = -0.201$, 95% CI [-0.392, -0.009]). See Figure 7.5.2.1, Panel E for a bar plot.

See Supplement S16 for a detailed summary table with variable descriptive statistics and differences scores.

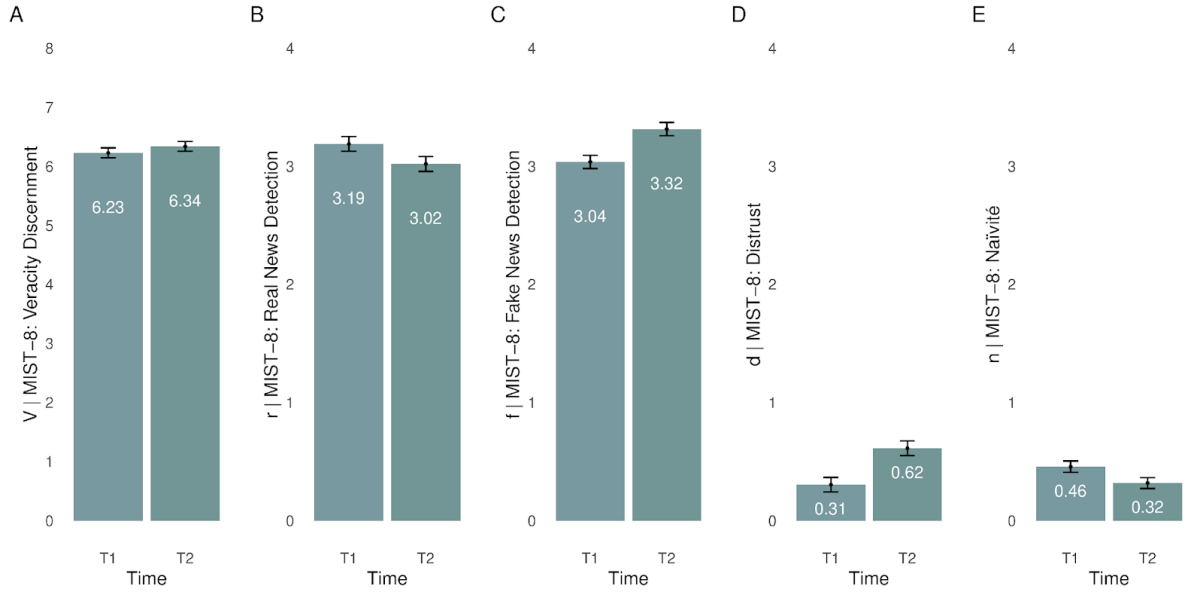

Figure 7.5.2.1. Plot of ***Verification done*** variables applied to the Bad News Game (N = 421). T1 = pre-test. T2 = post-test.

7.5.3 Discussion

Study 7C showed that using the MIST in conjunction with the ***Verification done*** framework provided novel insights contrary to our expectations—participants did not become better at general news veracity discernment after playing the *Bad News Game*. Looking at the MIST facet scales, we did find significant differences in both fake news detection and real news detection. More specifically, we observed that while people improved in the detection of fake news, they also became worse at the detection of real news. Looking further at response biases, we can also see that the *Bad News Game* might increase general distrust in news headlines, while also diminishing naïvité. At first sight, these results seem to indicate that the intervention does decrease people's susceptibility to fake news and reduces general naïvité, but at a potential cost of increased general distrust (hyper skepticism). Whether this means the intervention works depends on the aim: to decrease susceptibility to misinformation, or to increase the ability to accurately discern real news from fake news. The ***Verification done*** framework allows interventionists to start differentiating these important questions both theoretically as well as empirically.

In addition, as recommended by our framework, these results need to be interpreted in conjunction with the norm tables. The general sample that was recruited, was already highly media literate. The first quartile of the pre-test MIST scores was higher than the population average (veracity discernment: $1^{st}\ Quartile_{Population}$ = 50% accuracy, $1^{st}\ Quartile_{Sample}$ = 75% accuracy). Effects of the intervention might therefore be different with a more representative sample, or for people performing worse during the pre-test phase.

The results of this study come with two caveats. First, the MIST-8 was used instead of the MIST-20. As is common for short scales (Rammstedt et al., 2021; Thalmayer et al., 2011)—while maintaining high psychometric quality—the parsimonious MIST-8 is less precise and less reliable than the MIST-20. Since the MIST-20 only takes about 2 minutes to

complete, we recommend researchers to use the MIST-20 whenever possible. Second, while we were sufficiently powered to detect effect sizes similar to the original evaluation of the intervention (Roozenbeek & van der Linden, 2019a), we did not have sufficient power to detect smaller nuances (Anvari & Lakens, 2021; Funder & Ozer, 2019; Götz, Gosling, et al., 2021).

The results of this study indicate the importance of looking at misinformation susceptibility in a more holistic way. Applying the ***Verification done*** framework, we discovered key new theoretical dimensions that previous research had overlooked. Evaluators of this intervention, and other interventions, can now disentangle and accurately measure the five dimensions of misinformation susceptibility, thereby expanding our understanding of both the underlying mechanisms as well as the intervention's practical impact.

7.6 Discussion—Towards A Multifaceted Framework

7.6.1 A Standardised Measurement Instrument

We explained the necessity of a multifaceted measurement of misinformation susceptibility, and based on theoretical insights from previous research, developed the ***Verification done*** framework. Then, in three studies and six samples from two countries, we developed, validated, and applied the Misinformation Susceptibility Test (MIST): a holistic test which allows the assessment of *veracity discernment ability*, its facets *fake news detection ability* and *real news detection ability*, and judgement biases *distrust* and *naïvité*.

In Study 7A, we derived a development protocol, generated a set of fake news headlines using the GPT-2 neural network—an advanced language-based machine learning algorithm—and extracted a list of real news headlines from neutral and well-trusted sources. Through psychometric analysis using factor analysis and item response theory, we developed the MIST-8 and the MIST-20 tests.

In Study 7B, we recruited four nationally representative quota samples, two each for the US and the UK, from three different recruitment platforms, to conduct a rigorous validation procedure with the aim of maximizing the measure's fidelity and generalizability. First, confirmatory factor analyses consistently favored the higher-order structure and yielded satisfactory properties that suggest high validity and good reliability of both the MIST-8 and the MIST-20. Second, adopting a wide-net approach, we constructed an extensive nomological network. We found the MIST-8 and MIST-20 to be consistently highly correlated with, and similar to, various fact-check tests—thus signaling convergent validity—while being clearly distinct from the existing conspiracy mentality questionnaire (CMQ) and the bullshit receptivity scale (BSR), hence providing evidence for discriminant

validity. Moreover, we presented MIST-20 and MIST-8 norm tables for both the UK and the US, based on our large, nationally representative samples.

In the applied Study 7C, we demonstrated how ***Verification done*** and the MIST can be employed in naturalistic settings, in this case to evaluate the effects of a highly popular inoculation intervention. Employing the MIST to evaluate interventions, in combination with the norm tables, we were able to uncover new mechanisms behind a well-known media literacy intervention, the *Bad News Game* (Maertens et al., 2021; Roozenbeek & van der Linden, 2019a), and highlight both weaknesses and strengths of this intervention that were not detected using the classical methods. Moreover, for the first time, we were able to disentangle the five dimensions of misinformation susceptibility, finding unexpected changes in judgement biases as well as in real news detection, which can inspire further research and theoretical development.

Limitations and Future Research

While we firmly believe that the MIST marks a substantial methodological advance in the field of misinformation research (Bago et al., 2020; Batailler et al., 2021; Roozenbeek, Maertens, et al., 2021; Rosellini & Brown, 2021; Zickar, 2020), it is of course not without limitations. An inevitable challenge of doing any type of systematic and methodologically rigorous news headline research lies in the fact, that what might be real news at one point in time might be outdated at a later point in time. Therefore, similar to an IQ test, it may be necessary to regularly update the MIST over time.

Another related limitation concerns the inherent difficulty of the MIST's cross-cultural application. While we are greatly encouraged by our finding that the MIST appears to be an equally effective measure in the UK as it is in the US-American cultural context in which it was originally developed, cross-cultural translation poses a challenge. For obvious reasons, a simple and direct translation will not be sufficient. At the same time, while

trustworthy news sources from which real news items could be extracted can doubtlessly be identified in any language, for the time being, the GPT-2 (Radford et al., 2019)—the advanced language-based neural network algorithm that we employed to generate fake news items—has mainly been trained on English language corpora and is thus unable to produce comparable output in other languages (but see de Vries & Nissim., 2020; Guillou, 2020; for promising initial applications in Dutch, Italian, and Portuguese). It is our hope that this may change in the future, enabling the field to develop non-English adaptations of the MIST that will empower researchers around the globe to capture the complex and multi-faceted reality of misinformation spread—and resistance.

We can see many more avenues for future studies using ***Verification done*** and the MIST. One example is the implementation of the MIST in geo-psychological studies (Ebert et al., 2021; Ebert, Götz, et al., 2022; Ebert, Mewes, et al., 2022; Rentfrow, Jokela, et al., 2015; Rentfrow, Gosling, et al., 2013) to identify misinformation hotspots and covariates with national, regional, and local levels of misinformation susceptibility. Another strand of research may further deepen our conceptual understanding of media literacy. For example, in light of the current findings, it appears that veracity discernment may encompass both a comparatively stable, trait-like component, as well as a more malleable skill component. Future studies may more clearly identify this distinction and find ways on how to best use these insights to devise effective interventions that foster better detection of both fake news and real news and in turn ultimately lead to greater genuine veracity discernment. Finally, we identify six immediate use cases for the MIST: 1) to pre-screen participants for studies, 2) as a covariate to investigate subgroups (e.g., that are highly susceptible to misinformation), 3) as a control variable in a model, 4) to map geographical regions to identify misinformation susceptibility hotspots, 5) to identify brain regions linked to misinformation susceptibility, and 6) to evaluate interventions.

Conclusion

Researchers lack a unifying conceptualization of misinformation susceptibility and too often use unvalidated measures of misinformation susceptibility. We therefore developed a new overarching, unifying and multi-faceted interpretation framework (i.e., ***Verification done)*** and a new, thoroughly validated measurement instrument based on this framework (i.e., the Misinformation Susceptibility Test; MIST). The current paper acts as a blueprint of integrated theory and assessment development, and opens the door to standardized and comparative misinformation susceptibility research. Researchers as well as practitioners can now make a thorough evaluation of media literacy interventions by comparing MIST scores using the norm tables and the ***Verification done*** framework. Using our standardized and psychometrically validated instrument allows for a comprehensive evaluation, and also permits for holistic comparison studies and tables to be compiled reporting all five ***Verification done*** scores. Practitioners in turn can use these scores and comparisons to choose interventions that best fits their needs. ***Verification done*** and the MIST can be employed across a range of psychological disciplines, ranging from cognitive neuroscience to social and personality psychology, to reveal the psychological mechanisms behind susceptibility to misinformation or to test the outcome of interventions.

Chapter 8

General Discussion

This dissertation started with a general introduction into misinformation research, its gravity on today's society, and the advances in research on countering it (Compton et al., 2021; van der Linden, Roozenbeek, et al., 2021). I discussed inoculation theory—*"the grandparent theory of resistance to attitude change"* (Eagly & Chaiken, 1993, p. 561)—as one of the major psychological theories to provide insight into bolstering attitudes against unwanted attacks. Further, I expanded and adapted inoculation theory as a means to counter misinformation, across multiple modes and paradigms, and tackling relevant topics such as climate change misinformation. I noted that although inoculation has proven to be effective, its longevity remained unknown, and although the theory has been in development for over 60 years, no theoretical model on the mechanisms of longevity had been developed.

8.1 Theoretical Advancements: The Memory-Motivation Model of Inoculation

Although the literature on inoculation theory has seen a noticeable expansion in the past decade (Compton et al., 2021; Traberg et al., 2022), with innovations ranging from therapeutic inoculation (Compton, 2020) to gamified inoculation (Roozenbeek & van der Linden, 2019a), it has seen limited innovation with regard to the core mechanisms, its longevity, and the mechanisms of decay. To solve the issue of decay in inoculation theory, I first go back to the core mechanisms of the inoculation framework, and then move beyond the traditional model.

Inoculation theory was designed as a social psychological theory and has seen much of its further development within the communication sciences, but less so in cognitive psychology. Over the past decades, questions within the same original paradigm have remained prominent, such as the role of "threat" (Compton, 2021; Richards et al., 2017;

Richards & Banas, 2018), and "affect" (Compton et al., 2022) within inoculation interventions. Although important topics, major advances in our theoretical and psychological understanding of the drivers of the inoculation effect, and especially its long-term effectiveness, have been limited.

Inspired by early research on the potential role of memory in inoculation interventions (Pfau et al., 2005), I chose to explore if memory is an important mechanism within the inoculation process. Memory has long been studied in other areas of misinformation research, such as in the debunking literature (for an overview, see Swire & Ecker, 2018). However, with the exception of the study by Pfau et al. (2005), inoculation theorists had not yet tapped into the wealth of insights cognitive psychology and memory research can provide. Surveying the literature of inoculation (**Chapter 1–2**) made clear that the memory literature could help to shed a new light on the underlying mechanisms of inoculation decay. Building on this, I integrated the current insights from the cognitive psychology of memory with existing insights from the inoculation literature, and proposed a new memory-motivation model of inoculation (see Figure 2.2.1).

Through a series of five studies (**Study 1–5**), using three different inoculation paradigms, we can now assess the validity and generalisability of a memory-based inoculation theory and answer our main research and empirical questions. I indeed find that memory is one of the most dominant factors in the inoculation intervention success, as well as its longevity, and the studies also unveiled the expected decay curves for each of the interventions.

In **Study 1** (Maertens et al., 2020), I found that text-based, passive, therapeutic, issue-specific inoculation interventions are effective, replicating the results of previous research (van der Linden, 2017; Cook et al., 2017), and can remain fully effective for at least one week, while the effect of a consensus-message-based treatment without inoculation does

not. This indicates that even when inoculation is passive, a strong initial effect can be established. In **Study 2**, I expanded on this study to test the memory-motivation model using objective and subjective measures of the model's components, and included longer timeframes (T3 *Mdn* = 29 days). The effect after week was in line with the meta-analytic effect size of inoculation (d_{Study2} = 0.46 $d_{MetaAnalysis}$ = 0.43; Banas & Rains, 2010), and reduced by 39% after one month (d_{0days} = 0.46, d_{8days} = 0.38, d_{29days} = 0.28). These studies indicate that the inoculation effect remains intact for at least 1 month without any booster intervention with text-based inoculation messages on climate change, but that light decay takes place. This is in line with findings by a recent study by Ivanov et al. (2018), who reported significant decay after 6 weeks but not after 4 weeks, as well as the meta-analysis by Bana and Rains (2010), that reported decay to typically start to take place after 2 weeks. Meanwhile, I found the first evidence that a booster intervention—in this case a repetition of the original inoculation message—can boost the effect to prevent any decay from happening, in particular by boosting memory of the intervention (d_{0days} = 0.46, $d_{29days,NoBoost}$ = 0.28, $d_{29days,Boost}$ = 0.48). This finding is in line with Ivanov et al. (2018) who found a repeated inoculation message to be effective at lengthening the inoculation effect. Investigating the underlying mechanisms showed that motivational threat and issue involvement are predictors of the outcome variable, but that objective memory of the inoculation intervention was the strongest predictor of the inoculation effect over time. We also found that motivation had a positive influence on inoculation memory, in line with the predictions of the memory-motivation model. However, there was no evidence for an effect of the intervention on motivation. As we also found that motivation directly influenced the outcome variable, an important question for future research is whether we can improve our inoculation interventions in order to elicit more motivation (Compton, 2021, 2022).

In **Study 3** (Maertens et al., 2021), I investigated the same questions in a gamified, active, broad, prophylactic, inoculation paradigm. I found that repeated testing can serve as an inoculation booster, allowing for the inoculation to remain fully intact for up to 3 months, potentially due to memory strengthening that comes with testing. However, when not testing repeatedly, the effect was no longer significant after 2 months. The study also showed that these findings were not due to item or item ratio effects. In **Study 4**, I expanded on this study by removing the repeated testing confound and splitting the sample in groups per posttest time point, and adding the same set of questions about memory and motivation. I replicated the main effect of the *Bad News* game, with a larger effect than the typical effect size found in inoculation interventions ($d_{\text{Study4}} = 0.78$, $d_{\text{MetaAnalysis}} = 0.43$; Banas & Rains, 2010), but also found that when an immediate posttest is not included, the inoculation effect is no longer significant after 9 days, a faster effect decay than anticipated. When looking into the mechanisms, memory arises as the most dominant predictor, followed by motivational threat. Meanwhile, those high in memory did show an inoculation effect at 9 days and at 29 days, and a new and short version of the *Bad News* presented after 9 days worked well as a memory booster—although not enough to show a general inoculation effect after 29 days. A test of the memory-motivation model revealed new insights that are different from the climate change paradigm. In this paradigm, only memory was a significant predictor of the outcome measure. Motivation did not have an influence on the outcome measure, nor did the inoculation intervention have an effect on motivation. However, motivation on its own did have a significant effect on memory formation at T1. In line with the findings from the climate change paradigm, we found evidence for a full mediation of the inoculation effect through memory. These findings provide a second evidence base for the memory-motivation model, and evokes similar questions about the role of motivation in inoculation paradigms

(Compton, 2021, 2022): is it less important than previously thought, or did we fail to capture or move it appropriately?

In **Study 5**, I explored the same questions in a scalable, video-based, passive, broad, prophylactic form of inoculation. The study shows that both a short and long inoculation video can serve as an effective inoculation intervention, and that they are similarly effective, with an effect size similar to the meta-analytic effect size of inoculation interventions ($d_{\text{LongVideo}} = 0.53$, $d_{\text{ShortVideo}} = 0.44$, $d_{\text{MetaAnalysis}} = 0.43$; Banas & Rains, 2010). When no immediate posttest is included, we can see a quick inoculation effect decay after the intervention in the course of the first two weeks, paralleled by a similar decay curve for memory. However, when an immediate posttest is implemented, the video inoculation effect can remain effective for up to at least 29 days, with limited decay ($d_{\text{Exp3,0days}} = 0.30$, $d_{\text{Exp3,29days}} = 0.23$). When investigating the mechanisms, memory was again the most dominant predictor of the inoculation outcome. In addition, to disentangle the mechanisms further and explore the potential of booster interventions, we tested three booster videos, a repetition of the original video, a threat-focused booster (Booster A), and a technique memory booster (Booster B). Both the memory booster video and the repeated original inoculation video served successfully to boost inoculation memory, while the threat booster video did not impact memory performance and failed to increase motivation. Similar to Study 2 and Study 4, both memory and motivation were significant predictors of the inoculation effect. However, this time the original intervention did have a direct positive impact on motivation as well, which could mean that different types of inoculation interventions work through different mechanisms. Different from the previous two tests of the memory-motivation model, we now were able to disentangle T1 and T3 effects, and found that the intervention successfully improved memory *and* motivation at T1, and that motivation improved memory

at T1, which in turn increased the inoculation effect at T3 through memory at T3, in line with the memory-motivation model predictions.

The studies in this dissertation represent a first look at developing a memory paradigm for inoculation. Many of the choices made in this dissertation, including the choice of measures for memory and motivation, the causal ordering of the SEM models, and the used inoculation interventions, mean that the validity of the model needs to be thoroughly and independently tested before it can become the new standard. Nevertheless, taken together, these first direct measures of the mechanisms behind the longevity of the inoculation effect, explored across 5 studies (9 experiments) and 3 paradigms, suggest that a memory-motivation theory is a new feasible paradigm to consider and explore further. It also helps to formulate a data-driven answer to the main research question presented at the beginning of the dissertation: *"Can we explain the resistance to persuasion decay process using a memory-motivation theory of inoculation decay?"*. Not only did we find evidence for a role of memory in the explanation of the inoculation effects, the data suggests that it presents a better explanation of the longevity of the effects than the traditional account (Compton, 2013; McGuire, 1961a, 1961b, 1962, 1964, 1973; McGuire & Papageorgis, 1961, 1962). We do not find evidence for the need for an endured sense of threat for the inoculated persons to defend themselves against misinformation attacks at later time points, although we do find some evidence that motivation helps and that it could improve how much people learned from the inoculation intervention. In other words, these findings provide crucial new insights into resistance to persuasion. While threat and motivation have often been mentioned as a crucial aspect of inoculation, and proposed to be elicited as part of the "affective forewarning" in inoculation messages, literature studies show that the original authors and the first generation of inoculation scholars often did not manipulate nor measure it (Compton, 2013, 2021, 2022), although recently more research has been done that establishes its role

(Banas & Richards, 2017; Ivanov, 2012; Richards & Banas, 2018). In the text-based and gamified interventions we did not manage to manipulate motivational threat through our inoculation intervention, and therefore might have missed a crucial aspect of what constitutes an inoculation intervention, but we did manage to do so in the video-based intervention. All studies in this dissertation are congruent however in their finding that motivation can have a role, but that the role of memory is typically larger. Therefore, the data in this dissertation provides a strong basis for an alternative theory positing that inoculation effects are rooted in memory networks and can be trained accordingly (Pfau et al., 2005), opening up new possibilities by implementing insights from cognitive psychology related to learning, memory strengthening, and forgetting (Collins & Loftus, 1975; Ebbinghaus, 1885; Frankland & Bontempi, 2005; Hardt et al., 2013; Murre & Dros, 2015; Murty & Dickerson, 2016; Smith, 1998). The model proposed in Chapter 2 of this dissertation provides an example of such a model. Although promising, it has to be taken into account that this is a primitive first version of the model and needs to be replicated and tested in different forms in future research. It is for example possible that there are important aspects of inoculation interventions that moderate the relationships between the variables in the model. Some interventions may for example work by manipulating motivational threat in a different way (e.g., some inoculation interventions may work via memory, others via motivation, and others by manipulating both). There are also many alternative SEM models that could theoretically be viable instead of the currently used one, for example with a different causal ordering (e..g., it is possible that memory at T1 influences motivation at T1 instead of the other way around). Future research will need to disentangle these mechanisms and effects further.

Not only does the new memory-based approach present theoretical relevance, it also presents a practical benefit for intervention developers and policy makers, as we can now start to form an answer to the empirical question of this hypothesis: *"What is the shape of the*

inoculation effect decay curve and can booster interventions remediate the decay?". Plotting the data from all three paradigms together, we do indeed find a decay curve that resembles an exponential function, what one would expect when looking at an Ebbinghaus forgetting curve (Ebbinghaus, 1885; Murre & Dros, 2015). Figure 8.1.1 depicts the inoculation effect curve over time for each of the interventions in the first column, taken from Study 2, Study 4, and Study 5[98], and their respective memory forgetting curve in the second column. As can be seen, the decay curves of the inoculation effects are remarkably similar to the forgetting curves of the inoculation memory. When taking into account that the memory measures were newly created for each study and had not been validated before, and that each of the inoculation interventions had both very different modes of presentation (*text-based*, *gamified*, *video-based*) as well as very different outcome measures (*perceived scientific consensus*, *reliability rating of fake news*, and *discernment of manipulativeness*), this congruence is a promising first step towards unveiling the true decay curve of inoculation effects. In panel A and panel B, we can see that there is a strong inoculation effect for text-based inoculation as well as a strong memory, but that they, in a similar fashion, decay over time in what closely resembles an exponential curve, with some memory and some inoculation effect remaining after 29 days. When boosted almost the full inoculation effect together with the full inoculation memory remains. A very similar pattern can be found for the gamified intervention in panel C and panel D. Despite the effect no longer being significant after 9 days, we do see the congruence between the inoculation effect and the memory forgetting curve. Finally, in the video-based intervention, as seen in panel E and F, we again find that the memory forgetting curve, as well as the booster curve, is closely mirroring the inoculation effect pattern. In addition, we find that after 1–2 weeks, the effect is no longer significant, but

[98] For the plot of Study 5 I combined the datasets of the three experiments, taking the control group and single inoculation group from Experiment 1, and grouping all second posttests from the inoculation groups from Experiments 2–3 as booster groups due to repeated testing and the various booster interventions. To simplify the plot the time points were grouped based on the median days after T1 (i.e., T2 = 9 days, T3 = 29 days).

when boosted, the effect decays much slower than it decayed initially, and the same can be found for the inoculation memory. This shows what we would expect from a forgetting curve, namely, that once the memory is strengthened, the decay afterwards is slower (Ebbinghaus, 1885; Murre & Dros, 2015). The contrasts between the effectiveness of the interventions, and whether they are still significant after 1–2 weeks or not, could be explained by differing decay curves as well as by differing initial effectiveness. The results depicted here indicate that there is some uniformity in the decay curves across interventions, but also some variety in the rate of the decay. Whether this is a side effect of the different outcome measures or the small variations in the memory questions, or due to actual differences in effects or memory strength, will need to be explored in future research.

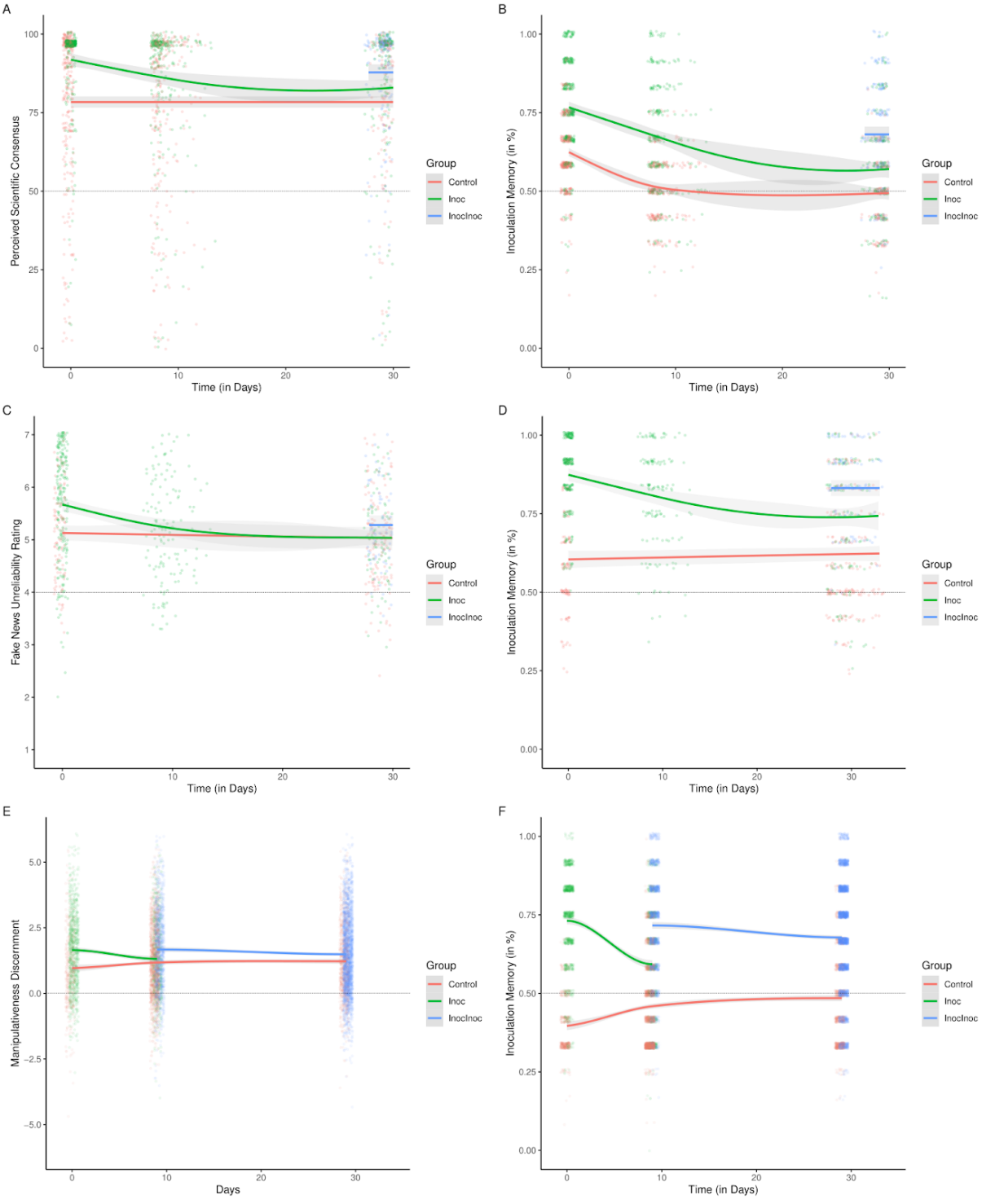

Figure 8.1.1. Line graphs of the inoculation effects (first column) and the objective inoculation memory (second columns) for each of the three intervention paradigms (A, B: text-based inoculation, $N_{datapoints}$ = 5,475; C, D: gamified inoculation, $N_{datapoints}$ = 2,022; E, F: video-based inoculation, $N_{datapoints}$ = 7,505). Inoc = single inoculation group. InocInoc = boosted inoculation group. Error bands represent the standard error.

In a recent literature review on the importance of "threat" in inoculation theory, inoculation scholar Compton (2021, p. 4299) posits that instead of moving away from the biomedical metaphor that inoculation theory is based on, we should rather focus on *"expanding our working understanding of medical inoculations, including checking in with recent developments in medical inoculation practices."* Our new memory theory of inoculation fits within this recommendation, as it builds upon the original metaphor: biomedical inoculation fosters the creation of antigen-specific memory T and B cells (cf. inoculation providing cognitive defences against misinformation techniques; Ciabattini et al., 2021; Jung et al., 2022). Similarly, for the idea of rehearsal and memory strengthening, we can refer to how most people vaccinated against COVID-19 still need a booster shot for full immunity—for vulnerable people, in some cases even a third booster shot is recommended to increase the longevity of the defences. Future research could therefore explore repeated inoculation booster shots, as it might show useful for true long-term effectiveness similar to the findings of Study 3 (Experiment 1) and Study 5 (Experiments 2–3).

Some inoculation scholars have argued that in the days after the inoculation intervention, the inoculation effect might in fact increase rather than decrease, as the inoculation effect might have to "sink in" (Dillingham & Ivanov, 2016; Ivanov, Pfau, et al., 2009; Ivanov, Parker, et al., 2018; McGuire, 1964). However, the results of this dissertation do not point in this direction, in fact, the evidence found for a potential exponential decay curve points toward the opposite: the decay is more likely to be higher in the first weeks. It is possible that this theory—which posited the benefits of an initial period of delay, but has limited empirical evidence (Banas & Rains, 2010)—came into existence due to a lack of high-powered studies systematically looking at the decay curve and the mechanisms of decay, and therefore the lack of general insights and understanding of the decay process. The findings from this dissertation seem to indicate that we might now have a feasible alternative

theory to fill this gap: the memory-motivation model may complement the traditional inoculation model that was based on threat, motivation, and counterarguing, by adding a memory dimension to explain its long-term effectiveness. In terms of the long-term effectiveness of inoculation, this represents a major innovation in explaining the core mechanisms of inoculation theory, and provides a blueprint for the benefits of bridging insights from subdisciplines within a field of specialisation.

8.2 Methodological Advancements: Longitudinal Designs, Testing Effects, and Multidimensional Psychometrics

The study of the long-term effectiveness of inoculation effects should not be limited to the decay of resistance to persuasion, as Hill et al. (2013, p. 542) stressed:

> *"Decay of the effects of political persuasion is too important to be ignored, as it routinely has been. It is a basic feature of mass persuasion in most if not all political contexts. Scholars should therefore try harder to build measurement of decay into their research designs."*

I would go further than this statement and stress that during the literature review of this dissertation, it became clear that important theories are often accepted with limited or no longitudinal research. This is not surprising: longitudinal research, especially with multiple timepoints and long-interval follow-ups, is both costly and difficult to run. A recent review of framing research highlights that long-term effects are often not measured, and that most longitudinal studies do not look beyond two weeks (Lecheler & de Vreese, 2016). Similarly, in a meta-analysis of behavioural intervention studies regarding action on climate change, Nisa et al. (2019, p. 9) stressed that they *"could not provide a definitive answer on persistent effects per specific type of intervention due to the small number of papers that reported follow-up effects"*. Within this dissertation I developed and tested various formats of different

longitudinal designs for each inoculation paradigm, with and without booster treatments. I looked beyond the standard single-treatment study and mapped long-term cognitive changes, thereby providing valuable insights relevant for the wider field of psychology, and inviting other researchers to consider a longitudinal design in their future studies as well.

During this journey, three important methodological questions—that were previously unanswered for inoculation research—were explored: item effects, testing effects, and psychometric validity. In **Study 6**, I looked at the potential confound of the inoculation effect interpretations caused by the use of a pretest and the researcher's choice of items as the dependent variable for misinformation reliability ratings. The research showed that there is no evidence for an effect caused by the implementation of a pretest, but that the specific choice of items has an influence on the effect size found. In this case the effect remained significant despite the change in items, but recent research by Roozenbeek, Traberg, et al. (2022) replicated the item effect and found that the effect can even change in direction when items are different for the pre and post tests. In other words, researchers have to be careful when choosing their items, and it stresses the need to work towards standardised item sets.

Despite there being no evidence for pretesting effects, the other studies in this dissertation made clear that there are two other testing effects that we do need to take into account: the immediate posttest and repeated posttests. The small difference in design between **Study 3** (with immediate posttest) and **Study 4** (without immediate posttest), and between **Study 5, Experiment 1** (without immediate posttest) and **Study 5, Experiments 2–3** (with immediate posttest), demonstrated that the use of an immediate posttest potentially serves as an immediate memory booster. This is an important finding both for intervention implementation guidelines and for intervention evaluation science. It shows that an immediate posttest may not be advisable for evaluations of the long-term effectiveness evaluation of an intervention. While it could be argued that participants learn how to respond

to particular items, we did not find evidence for this in **Study 5**, where participants had to discern the manipulativeness of a random set of headlines from a larger pool of social media posts at each time point, with the possibility that items of the same topic switch from manipulative to neutral between time points. Similar to the immediate posttest effect, the difference in design between **Study 3, Experiment 1** (with immediate posttest *and* repeated posttests at multiple time points) and **Study 3, Experiment 2** (with immediate posttest but *no* additional repeated posttests until the final time point), shows that repeating a posttest at multiple time points may serve as an additional booster on top of the immediate posttest. Also this finding fits into findings from the literature outside of the inoculation scholarship, in particular from cognitive psychology, with previous research finding similar learning effects by repeated testing (Karpicke & Roediger, 2008; Linton, 1975; Roediger & Karpicke, 2006a, 2006b). Combined, the immediate posttest and the repeated posttest effects indicate that one should ideally use a design that exposes each participant to a maximum of one posttest (e.g., with each participant or group of participants receiving the posttest at a different point in time after the intervention, similar to **Study 2**, **Study 4**, and **Study 5, Experiment 1**). This finding also has a positive side—it indicates that if practitioners are implementing an intervention in the field, it may be useful to consider including a quiz or a feedback mechanism at the end of the intervention to consolidate participants' knowledge, and repeatedly follow-up with the participants of the intervention over time, to further strengthen and increase the longevity of the effects.

The item effects found in **Study 6** on the other hand demonstrated that to study inoculation effects, next to the measurement over time and the choice of an adequate research design for it, the measurement instrument used is also of importance. Moreover, with the proliferation of research on misinformation, came a surge in unstandardised research methodologies and unvalidated scales (van der Linden, 2022). If we do not understand the

structure of misinformation susceptibility and do not have the psychometric toolkit to measure it, too much error variance obscures the true results behind the multitude of misinformation studies. In addition to the new memory theory, and the new insights on longitudinal designs to measure inoculation strength over time, I therefore reflected on the importance of accurate measurement tools. In **Study 7**, I worked towards a new measurement framework as well as a measurement instrument to provide more precise insights into misinformation susceptibility. The Misinformation Susceptibility Test (MIST), as well as the new ***Verification done*** framework that was created, could be considered the first psychometrically validated measurement instrument and interpretation schema for misinformation susceptibility (Maertens et al., 2022). Moreover, the process of creating the new toolkit was supported by the implementation of innovative methods in artificial intelligence for representative item generation (Götz, Maertens, et al., 2021), and multidimensional item-response theory for the mapping of the hierarchical discernment ability. Future studies will be able to make use of the MIST either as a covariate or as an outcome variable, as well as the ***Verification done*** framework in the context of intervention design and testing.

In the beginning of the dissertation I opened the methodological question on the backbone of this work; *"Is our methodological toolkit adequate to accurately measure the long-term effectiveness of inoculation effects?"* Exploring this question made clear there are many potential issues to take into account, from testing and item effects, to validated measurement instruments. By providing a wide range of methodologies in this dissertation, while not providing a definite answer, I am confident that we are getting closer to robust methodological approaches. Across the studies, I have varied the research designs by excluding or including pretests (see Solomon, 1949, and Gelman, 2017, for discussions of pretesting effects and how to take them into account), immediate posttests, and repeated

posttests. I have varied the dependent variable by including indicated belief (Studies 1–2), misinformation reliability ratings (Studies 3–4), and manipulativeness discernment (Study 5) as outcome measures. I have varied item ratios from equal to unequal (see Aird et al., 2018, for an example of ratio effects in a different context), and item sets from the same to different. An additional challenge for the field is that—although question framings and response modes do not seem to have much influence on the outcome (Roozenbeek, Maertens, et al., 2022)—researchers often use random headlines from the internet and use these as their scales without psychometric standardisation or without considering the proportion of manipulation techniques used (see e.g., Pennycook, Binnendyk, et al., 2021), which makes it difficult to make items comparable across interventions and could provide additional confounds with manipulative technique ratios, measurement error, social cues, source cues, memory, and content-specific effects (Traberg et al., 2022; van der Linden, Roozenbeek et al., 2021; van der linden, 2022). The MIST, or similar tools, could be further developed and employed in studies to avoid some of these issues. In addition, the standardised nature of the MIST could make it useful for new applications beyond the scope of this dissertation, such as comparing susceptibility across regions, mapping misinformation susceptibility hotspots in countries, and predicting regional vaccine uptake based on misinformation susceptibility data (see e.g., Ebert, Götz, et al., 2022; Ebert, Mewes, et al., 2022; Rentfrow, Jokela, et al., 2015; Rentfrow, Gosling, et al., 2013). Meanwhile, even when not using the MIST, researchers and practitioners could start implementing the *Verification done* framework to disentangle real news and fake news effects, and extract a true underlying discernment ability, as well as response biases.

Another concrete example of how using a variety of methods can reveal unexpected patterns, is when looking at the effect of inoculation interventions on evaluations of trustworthy news and general scepticism. In Study 3 (Experiments 1–2), we found that the

Bad News game could be reducing the reliability evaluations of real news items, which would be an undesired effect. However, when the item set was changed and a delay introduced, this effect disappeared (Study 3, Experiment 3), leading to the conclusion that this real news effect is negligible and mostly due to measurement error. In Study 6 we found the same initial result, but when standardising the items and removing the pretest, this effect again disappeared, leading us once more to a similar conclusion. However, in Study 7, using a more rigorous measurement test (the MIST-8), we again found that people tend to become more sceptical of real news after the intervention—although here it has to be noted that the *Bad News* intervention focuses on *manipulative* (misleading/partially false) content while the MIST uses *true* vs *false* statements. While we did not explore whether the result would replicate with a delay or without a pretest, the consistent finding that there is an effect on general scepticism across measurement instruments does provide some evidence that this is not just measurement error. Future research should further unpack this, and also ask the question whether we can develop inoculation interventions that promote building trust in real news as well as helping people detect misinformation.

The work on methodology that forms the backbone of this dissertation made it clear that theoretical and empirical advancements go hand in hand with methodological advancements, unveiling a plethora of new insights that would not have been possible by focusing just theory. I hope that it will inspire other researchers to focus on and give equal weight to all three of these aspects (*theory*, *data*, and *methodology*).

8.3 Impact and Applications: Towards Better Interventions and Measurement

The theoretical and methodological advancements provided by the dissertation have various implications, including for applied interventions. In this section I zoom out and provide some insights on the "big-picture" implications of this work.

In this work I showed that memory has a role in inoculation theory effectiveness, as well as its longevity, and that it is a better predictor of the long-term effectiveness of such interventions than other mechanisms such as motivation, but that in a best-case scenario, both are combined. When researchers or practitioners want to predict the inoculation effect decay curve (i.e., the long-term effectiveness) this can be best done by using an exponential forgetting curve to model the effects. Similarly, when trying to establish the remaining cognitive immunity after letting participants go through an intervention, this would be best done by including a measure of memory. However, motivation remains important to establish and maintain a good memory of the intervention, and should therefore still be taken into account in intervention design. Some of these insights could potentially be extended to other theories on (resistance to) persuasion, as well as other social psychological theories. This solves a debate in recent literature (e.g., Maertens et al., 2021) on whether memory and cognitive psychology can be integrated with a social psychological theory that focuses more on threat and motivation. It shows that while memory is a better predictor of inoculation effects, it is productive to integrate the various mechanisms into a coherent joint model. Researchers and practitioners should always consider what insights are available in the various subdisciplines of their field as well as beyond, to look beyond their own perspective and be open to integrating existing findings that can be generalised across disciplines.

The insights on the decay of the intervention effects also show that it would be beneficial if practitioners implemented booster interventions relatively close to the original intervention (within the first days or weeks), rather than later, as there is strong evidence that the decay is following an exponential pattern where most of the inoculation memory is lost at an early stage. This seems to be relevant regardless of the type of inoculation intervention (e.g., video-based or gamified): although the decay rate may differ, the decay function is similar across interventions. The insight into memory as a core mechanism of inoculation

effects also leads to various implications on how to best design inoculation interventions and slow down that decay. For example, researchers and practitioners may want to make sure the interventions are as memorable as possible. One could include repetitions of the important content in the same message to immediately strengthen memory, and make sure that the inoculation content is coherent and salient, so that it can easily be captured and stored in memory, as well as easily retrieved (cf. debunking handbook; Lewandowsky, Cook, et al., 2020). Finally, it is important to make sure that any later booster intervention clearly reiterates the points learned previously. The methodological findings further show that it is useful to take into account any boosting effects by testing participants regularly, or even to specifically harness the power of testing participants (when possible with feedback) to increase learning.

Finally, researchers and practitioners are encouraged to look at inoculation effects holistically, as regardless of whether the inoculation effect or memory is strong or weak, the research shows that the intervention can differentially influence a range of relevant variables, such as scepticism, response biases, trust, and discernment. Similarly, the research shows that it is good practice to compare intervention participants with the general population, also in terms of misinformation susceptibility, to make better judgements on the effects on specific subgroups of the population. It is important researchers consider comprehensive and standardised measurement toolkits to do so, such as the Misinformation Susceptibility Test—which is now freely available to the research community (Maertens et al., 2022)—that give multiple resilience scores and have been explored in relation to a wide range of secondary variables. The use of tools such as the ***Verification done*** framework presented in the dissertation could also benefit practitioners as it presents an easy-to-use, interpretable, yet comprehensive overview of how discernment, specific detection skills, and biases each play a role in someone's skill level and how an interventions influence each of these variables.

In summary, with this dissertation I show the benefits of an open-minded and cross-disciplinary approach to theory development both in inoculation research and beyond, and I hope that the concrete suggestions above will lead to better applied interventions and improved design and evaluation decisions made by both researchers and practitioners.

8.4 Conclusion

The series of studies presented in this dissertation provide a response to three important theoretical, empirical, and methodological questions: 1) what are the mechanisms behind the inoculation effect and its longevity, 2) what does the inoculation effect decay curve look like, and 3) how can we accurately measure misinformation susceptibility and the long-term effectiveness of inoculation effects? By integrating insights from cognitive psychology with those found in social psychology, I designed a new memory-motivation theory on inoculation. In a series of experiments, I then unveiled the inoculation decay function and explored the importance of memory of the inoculation intervention. Additional evidence pointed towards motivation as a potential memory enhancer, thereby bridging the memory theory with the threat-motivation theory. In addition, I illuminated the underlying theoretical mechanisms of memory strengthening, finding that a regular booster treatment may be needed to enhance the inoculation effect by strengthening memory of the intervention. A comparison across three intrinsically different paradigms (text-based, gamified, and video-based), each utilising different inoculation parameters, allowed us to indicate the validity and generalisability of the memory-motivation theory of inoculation. Strong evidence for the important role of memory, and a model combining motivation and memory performed well across all paradigms. This new evidence for the dominant role of memory in the longevity of inoculation effects provides a major advancement in inoculation theory, potentially shifting away attention from threat and motivation, which has been an important focus for the past 60 years of inoculation research (Compton 2021, 2022). I also

found that study design choices are an important parameter in effect evaluation, explaining the presence of both testing and item effects. I discussed that unstandardized measurement methods could mislead the interpretation of misinformation research, and that therefore standardised frameworks should be encouraged. In line with this, I developed the first psychometrically validated test and interpretation framework for the measurement of misinformation susceptibility: the MIST and ***Verification done***, which resemble an important innovation in methodology within the misinformation scholarship. In addition to presenting the new memory model of inoculation, the findings and methodologies used in this dissertation are insightful in light of broader issues related to the decay of effects within psychology, and the standardisation of scales and measurement frameworks, aiming to inspire a new stream of longitudinal, integrative, and robust research across the different branches of psychology and beyond.

References

Aichholzer, J., & Kritzinger, S. (2016). Kurzskala politischer Zynismus (KPZ). [Short scale of political cynicism]. *Zusammenstellung Sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis245

Aird, M. J., Ecker, U. K. H., Swire, B., Berinsky, A. J., & Lewandowsky, S. (2018). Does truth matter to voters? The effects of correcting political misinformation in an Australian sample. *Royal Society Open Science*, *5*(12), Article 180593. https://doi.org/10.1098/rsos.180593

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *The Journal of Economic Perspectives: A Journal of the American Economic Association*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

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). https://doi.org/10.1126/sciadv.aay3539

Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., Dasgupta, P., & Dubash, N. K. (2014). *IPCC fifth assessment synthesis report-climate change 2014 synthesis report*. https://www.ipcc.ch/site/assets/uploads/2018/02/SYR_AR5_FINAL_full.pdf

Amazeen, M. A. (2020). Journalistic interventions: The structural factors affecting the global emergence of fact-checking. *Journalism*, *21*(1), 95–111. https://doi.org/10.1177/1464884917730217

Anderegg, W. R. L., Prall, J. W., Harold, J., & Schneider, S. H. (2010). Expert credibility in climate change. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(27), 12107–12109. https://doi.org/10.1073/pnas.1003187107

Anderson, J. R. (1983). *Cognitive science series. The architecture of cognition*. Lawrence Erlbaum Associates, Inc.

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*. Advance online publication. https://doi.org/10.1016/j.jesp.2021.104159

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Arun, C. (2019). *On WhatsApp, rumours, lynchings, and the Indian Government*. https://papers.ssrn.com/abstract=3336127

Atkins, P. W. B., & Murre, J. M. J. (1998). Recovery of unrehearsed items in connectionist models. *Connection Science*, *10*(2), 99–119. https://doi.org/10.1080/095400998116521

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, *149*(8), 1608–1613. https://doi.org/10.1037/xge0000729

Bakir, V., & McStay, A. (2018). Fake news and the economy of emotions: Problems, causes, solutions. *Digital Journalism*, *6*(2), 154–175. https://doi.org/10.1080/21670811.2017.1345645

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

Banas, J. A., & Richards, A. S. (2017). Apprehension or motivation to defend attitudes? Exploring the underlying threat mechanism in inoculation-induced resistance to

persuasion. *Communication Monographs*, *84*, 164–178. https://doi.org/10.1080/03637751.2017.1307999

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science*, *26*(10), 1531–1542. https://doi.org/10.1177/0956797615594620

Baron, J. (2019). Actively open-minded thinking in politics. *Cognition*, *188*, 8–18. https://doi.org/10.1016/j.cognition.2018.10.004

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about Bad News: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1), 2. https://doi.org/10.5334/joc.91

Basol, M., Roozenbeek, J., McClanahan, P., Berriche, M., Uenal, F., & 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/10.1177/20539517211013868

Batailler, C., Brannon, S. M., Teas, P. E., & Gawronski, B. (2020). A signal detection approach to understanding the identification of fake news. *Perspectives on Psychological Science*. In press. https://www.researchgate.net/publication/344693250_A_Signal_Detection_Approach_to_Understanding_the_Identification_of_Fake_News

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Berman, M. G., Jonides, J., & Lewis, R. L. (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(2), 317–333. https://doi.org/10.1037/a0014873

Biermann, F., & Boas, I. (2010). Preparing for a warmer world: Towards a global governance system to protect climate refugees. *Global Environmental Politics*, *10*(1), 60–88. https://doi.org/10.1162/glep.2010.10.1.60

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Bouchet-Valat, M. (2020). Package 'SnowballC'. *The Comprehensive R Archive Network*. https://cran.r-project.org/package=SnowballC

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1), 7. https://doi.org/10.1038/s41467-018-07761-2

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 1745691620917336. https://doi.org/10.1177/1745691620917336

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brick, C., Hood, B., Ekroll, V., & de-Wit, L. (2021). Illusory essences: A bias holding back theorizing in psychological science. *Perspectives on Psychological Science*. https://doi.org/10.31234/osf.io/eqma4

Bridgewater, P., Loyau, A., & Schmeller, D. S. (2019). The seventh plenary of the intergovernmental platform for biodiversity and ecosystem services (IPBES-7): A global assessment and a reshaping of IPBES. In *Biodiversity and Conservation*. https://doi.org/10.1007/s10531-019-01804-w

Bronstein, M. V., Pennycook, G., Bear, A., Rand, D. G., & Cannon, T. D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition*, *8*(1), 108–117. https://doi.org/10.1016/j.jarmac.2018.09.005

Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in Psychology*, *4*, 1–15. https://doi.org/10.3389/fpsyg.2013.00279

Brown, G. D. A., & Lewandowsky, S. (2010). Forgetting in memory models: Arguments against trace decay and consolidation failure. In S. Della Sala (Ed.), *Forgetting* (pp. 49–75). Psychology Press. https://doi.org/10.4324/9780203851647

Brown, J. (1958). Some tests of the decay theory of immediate memory. *The Quarterly Journal of Experimental Psychology, 10*(1)*,* 12–21. https://doi.org/10.1080/17470215808416249

Bruder, M., Haffke, P., Neave, N., Nouripanah, N., & Imhoff, R. (2013). Measuring individual differences in generic beliefs in conspiracy theories across cultures: Conspiracy mentality questionnaire. *Frontiers in Psychology*, *4*(279), 1–15. https://doi.org/10.3389/fpsyg.2013.00225

Brydges, C. R., Gignac, G. E., & Ecker, U. K. (2018). Working memory capacity, short-term memory capacity, and the continued influence effect: A latent-variable analysis. *Intelligence*, *69*, 117–122. https://doi.org/10.1016/j.intell.2018.03.009

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, *114*(3), 542–551. https://doi.org/10.1037/0033-2909.114.3.542

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *13*(2), 149–154. https://doi.org/10.1177/1745691617706516

Camina, E., & Güell, F. (2017). The neuroanatomical, neurophysiological and psychological basis of memory: Current models and their origins. *Frontiers in Pharmacology*, *8*, Article 438. https://doi.org/10.3389/fphar.2017.00438

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://www.ncbi.nlm.nih.gov/pubmed/13634291

Carlton, J. S., Perry-Hill, R., Huber, M., & Prokopy, L. S. (2015). The climate change consensus extends beyond climate scientists. *Environmental Research Letters*, *10*(9), 094025. https://doi.org/10.1088/1748-9326/10/9/094025

Carpenter, J., Preotiuc-Pietro, D., Clark, J., Flekova, L., Smith, L., Kern, M. L., Buffone, A., Ungar, L., & Seligman, M. (2018). The impact of actively open-minded thinking on social media communication. *Judgment and Decision Making*, *13*(6), 562–574. http://journal.sjdm.org/18/18328/jdm18328.pdf

Carpenter, S. (2018). Ten steps in scale development and reporting: A guide for researchers. *Communication Methods and Measures*, *12*(1), 25–44. https://doi.org/10.1080/19312458.2017.1396583

Carrasco-Farré, C. (2022). The fingerprints of misinformation: How deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions. *Humanities and Social Sciences Communications*, *9*(1), 1–18. https://doi.org/10.1057/s41599-022-01174-9

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the Renvironment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., ... & De Rosario, M. H. (2021). pwr: Basic functions for power analysis. *The Comprehensive R Archive Network*. https://cran.r-project.org/package=pwr

Chan, K., Dupuy, B., & Lajka, A. (2020). *Conspiracy theorists burn 5G towers claiming link to virus*. ABC News. https://abcnews.go.com/Health/wireStory/conspiracy-theorists-burn-5g-towers-claiming-link-virus-70258811

Chan, M.-P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Chiluwa, I. E. (2019). Deception in online terrorist propaganda: A study of ISIS and Boko Haram. In I. E. Chiluwa & S. A. Samoilenko (Eds.), *Handbook of research on deception, fake news, and misinformation online* (pp. 520–537). IGI Global. https://doi.org/10.4018/978-1-5225-8535-0.ch028

Chołoniewski, J., Sienkiewicz, J., Dretnik, N., Leban, G., Thelwall, M., & Hołyst, J. A. (2020). A calibrated measure to compare fluctuations of different entities across

timescales. *Scientific Reports*, *10*(1), Article 20673. https://doi.org/10.1038/s41598-020-77660-4

Ciabattini, A., Pastore, G., Fiorino, F., Polvere, J., Lucchesi, S., Pettini, E., ... & Medaglini, D. (2021). Evidence of SARS-CoV-2-specific memory B cells six months after vaccination with the BNT162b2 mRNA vaccine. *Frontiers in Immunology*, *12*, Article 740708. https://doi.org/10.3389/fimmu.2021.740708

Cialdini, R. B. (1993). *Influence: The psychology of persuasion*. Harper Business.

Cialdini, R. B., Demaine, L. J., Sagarin, B. J., Barrett, D. W., Rhoads, K., & Winter, P. L. (2006). Managing social norms for persuasive impact. *Social Influence*, *1*(1), 3–15. https://doi.org/10.1080/15534510500181459

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(9), Article e2023301118. https://doi.org/10.1073/pnas.2023301118

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://doi.org/10.1037//1040-3590.7.3.309

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25–47. http://journal.sjdm.org/11/11808/jdm11808.pdf

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Compton, J. (2009). Threat explication . *STAM Journal*, *39*, 1–18.

Compton, J. (2013). *Inoculation theory.* In J. P. Dillard & L. Shen (Eds.), *The SAGE handbook of persuasion: Developments in theory and practice* (p. 220–236). Sage Publications, Inc. http://doi.org/10.4135/9781452218410.n14

Compton, J. (2013). Inoculation theory. *The Sage Handbook of Persuasion: Developments in Theory and Practice*, *2*, 220–237. https://doi.org/10.4135/9781452218410

Compton, J. (2019). Prophylactic versus therapeutic inoculation treatments for resistance to influence. *Communication Theory*. https://doi.org/10.1093/ct/qtz004

Compton, J. (2021). Threat and/in Inoculation Theory. *International Journal of Communication*, *15*, 4294–4306. https://ijoc.org/index.php/ijoc/article/view/17634

Compton, J., & Ivanov, B. (2012). Untangling threat during inoculation-conferred resistance to influence. *Communication Reports*, *25*(1), 1–18. https://doi.org/10.1080/08934215.2012.661018

Compton, J., Ivanov, B., & Hester, E. (2022). Inoculation Theory and Affect. *International Journal of Communication Systems*, *16*, 3470–3483. https://ijoc.org/index.php/ijoc/article/view/19094

Compton, J., & Pfau, M. (2005). Inoculation theory of resistance to influence at maturity: Recent progress in theory development and application and suggestions for future research. *Annals of the International Communication Association*, *29*(1), 97–146. https://doi.org/10.1080/23808985.2005.11679045

Compton, J., van der Linden, S., Cook, J. and Basol, M. (2021). Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories. *Social and Personality Psychology Compass*, *15*(6), Article e12602. https://doi.org/10.1111/spc3.12602

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Erlbaum Associates. https://www.routledge.com/A-First-Course-in-Factor-Analysis/Comrey-Lee/p/book/9781138965454

Cook, J., Lewandowsky, S., & Ecker, U. K. H. (2017). Neutralizing misinformation through inoculation: Exposing misleading argumentation techniques reduces their influence. *PloS One*, *12*(5), Article e0175799. https://doi.org/10.1371/journal.pone.0175799

Cook, J., Maibach, E., van der Linden, S., & Lewandowsky, S. (2018). *The consensus handbook*. https://doi.org/10.13021/G8MM6P

Cook, J., Nuccitelli, D., Green, S. A., Richardson, M., Winkler, B., Painting, R., Way, R., Jacobs, P., & Skuce, A. (2013). Quantifying the consensus on anthropogenic global warming in the scientific literature. *Environmental Research Letters*, *8*(2), 024024. https://doi.org/10.1088/1748-9326/8/2/024024

Cook, J., Oreskes, N., Doran, P. T., Anderegg, W. R. L., Verheggen, B., Maibach, E. W., Stuart Carlton, J., Lewandowsky, S., Skuce, A. G., Green, S. A., Nuccitelli, D., Jacobs, P., Richardson, M., Winkler, B., Painting, R., & Rice, K. (2016). Consensus on consensus: A synthesis of consensus estimates on human-caused global warming. *Environmental Research Letters*, *11*(4), 048002. https://doi.org/10.1088/1748-9326/11/4/048002

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, *7*(3), 613–628. https://doi.org/10.1017/psrm.2018.10

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation*, *10*(1), Article 7. https://doi.org/10.7275/jyj1-4868

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

Culliford, E. (2020). Facebook, YouTube remove 'Plandemic' video with 'unsubstantiated' coronavirus claims. *Reuters*. https://www.reuters.com/article/us-health-coronavirus-tech-video-idUSKBN22K077

Curley, A. (2020, September 18). *How to use GPT-2 in Google Colab*. The Startup. https://medium.com/swlh/how-to-use-gpt-2-in-google-colab-de44f59199c1

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. B. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods & Research*, *32*(2), 208–252. https://doi.org/10.1177/0049124103256130

De keersmaecker, J., & Roets, A. (2017). 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence*, *65*, 107–110. https://doi.org/10.1016/j.intell.2017.10.005

de Vries, W., & Nissim, M. (2020). *As good as new: How to successfully recycle English GPT-2 to make models for other languages*. ArXiv. https://arxiv.org/abs/2012.05628

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, *130*(6), 959–988. https://doi.org/10.1037/0033-2909.130.6.959

Dillingham, L. L., & Ivanov, B. (2016). Using postinoculation talk to strengthen generated resistance. *Communication Research Reports*, *33*(4), 295–302. https://doi.org/10.1080/08824096.2016.1224161

Doran, P. T., & Zimmerman, M. K. (2009). Examining the scientific consensus on climate change. *Eos, Transactions American Geophysical Union*, *90*(3), 22. https://doi.org/10.1029/2009EO030002

Drummond, C., & Fischhoff, B. (2017). Individuals with greater science literacy and education have more polarized beliefs on controversial science topics. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(36), 9587–9592. https://doi.org/10.1073/pnas.1704882114

Dunlap, R. E., McCright, A. M., & Yarosh, J. H. (2016). The political divide on climate change: Partisan polarization widens in the US. *Environment: Science and Policy for Sustainable Development*, *58*(5), 4–23. https://doi.org/10.1080/00139157.2016.1208995

Dür, A., & Schlipphak, B. (2021). Elite cueing and attitudes towards trade agreements: The case of TTIP. *European Political Science Review*, *13*(1), 41–57. https://doi.org/10.1017/S175577392000034X

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Harcourt Brace Jovanovich College Publishers.

Ebbinghaus, H. (1885). *Über das gedächtnis: Untersuchungen zur experimentellen psychologie*. Duncker & Humblot.

Ebert, T., Götz, F. M., Gladstone, J. J., Müller, S. R., & Matz, S. C. (2021). Spending reflects not only who we are but also who we are around: The joint effects of individual and geographic personality on consumption. *Journal of Personality and Social Psychology*, *121*(2), 378. https://doi.org/10.1037/pspp0000344

Ebert, T., Götz, F. M., Mewes, L., & Rentfrow, P. J. (2022). Spatial analysis for psychologists: How to use individual-level data for research at the geographically aggregated level. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000493

Ebert, T., Mewes, L., Götz, F. M., & Brenner, T. (2022). Effective maps, easily done: visualizing geo-psychological differences using distance weights. *Advances in*

Methods and Practices in Psychological Science, *5*(3). https://doi.org/10.1177/25152459221101816

Ecker, U. K. H., & Ang, L. C. (2019). Political attitudes and the processing of misinformation corrections. *Political Psychology*, *40*(2), 241–260. https://doi.org/10.1111/pops.12494

Ecker, U. K. H., Lewandowsky, S., & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition*, *38*(8), 1087–1100. https://doi.org/10.3758/MC.38.8.1087

Ecker, U. K. H., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology. Applied*, *20*(4), 323–335. https://doi.org/10.1037/xap0000028

Ecker, U. K. H., Lewandowsky, S., Fenton, O., & Martin, K. (2014). Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation. *Memory & Cognition*, *42*(2), 292–304. https://doi.org/10.3758/s13421-013-0358-x

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Farrell, J. (2019). The growth of climate change misinformation in US philanthropy: evidence from natural language processing. *Environmental Research Letters: ERL [Web Site]*, *14*(3), 034013. https://doi.org/10.1088/1748-9326/aaf939

Farrell, J., McConnell, K., & Brulle, R. (2019). Evidence-based strategies to combat scientific misinformation. *Nature Climate Change*, *9*, 191–195. https://doi.org/10.1038/s41558-018-0368-6

Fazio, L. K. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, *1*(2), 1–8. https://doi.org/10.37016/mr-2020-009

Fellows, I., Fellows, M. I., Rcpp, L., & Rcpp, L. (2018). Package 'wordcloud'. *The Comprehensive R Archive Network*. https://cran.r-project.org/package=wordcloud

Finch, J. F., & West, S. G. (1997). The investigation of personality structure: Statistical models. *Journal of Research in Personality*, *31*(4), 439–485. https://doi.org/10.1006/jrpe.1997.2194

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. *Political Psychology*, *38*, 127–150. https://doi.org/10.1111/pops.12394

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology*, *39*(2), 291–314. https://doi.org/10.1111/j.1744-6570.1986.tb00583.x

Forgas, J. P. (2001). *Feeling and thinking: The role of affect in social cognition*. Cambridge University Press.

Frankland, P. W., & Bontempi, B. (2005). The organization of recent and remote memories. *Nature Reviews Neuroscience*, *6*(2), 119-130. https://doi.org/10.1038/nrn1607

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives: A Journal of the American Economic Association*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Freedman, J. L., & Sears, D. O. (1965). Warning, distraction, and resistance to influence. *Journal of Personality and Social Psychology*, *1*, 262–266. https://doi.org/10.1037/h0021872

Frenkel, S., Alba, D., & Zhong, R. (2020). Surge of virus misinformation stumps Facebook and Twitter. *The New York Times*. https://www.bridgeportedu.net/cms/lib/CT02210097/Centricity/Domain/3754/Costello_Journalism_11_3_23.3_31.pdf

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Garrett, R. K. (2017). The "echo chamber" distraction: Disinformation campaigns are the problem, not audience fragmentation. *Journal of Applied Research in Memory and Cognition*, *6*(4), 370–376. https://doi.org/10.1016/j.jarmac.2017.09.011

Garry, A., Walther, S., Rukaya, R., & Mohammed, A. (2021). QAnon conspiracy theory: Examining its evolution and mechanisms of radicalization. *Journal for Deradicalization*, *26*, 152–216. https://journals.sfu.ca/jd/index.php/jd/article/view/437/265

Gelman, A. (2017, November 25). Poisoning the well with a within-person design? What's the risk? *Statistical Modeling, Causal Inference, and Social Science*. https://statmodeling.stat.columbia.edu/2017/11/25/poisoning-well-within-person-design-whats-risk/

Giner-Sorolla, R. (2018). *Powering your interaction*. https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

Goering, L. (2019). "Cranky Uncle" game offers a vaccination against climate disinformation. *Reuters*. https://www.reuters.com/article/us-climate-change-denial-app-idUSKBN1Y92F7

Goldberg, M. H., van der Linden, S., Ballew, M. T., Rosenthal, S. A., & Leiserowitz, A. A. (2019). The role of anchoring in judgments about expert consensus. *Journal of Applied Social Psychology*, *49*(3), 192–200. https://doi.org/10.1111/jasp.12576

Götz, F. M., Gosling, S. D., & Rentfrow, P. J. (2021). Small effects: The indispensable foundation for a cumulative psychological science. *Perspectives on Psychological Science*, 1–11. https://doi.org/10.1177/1745691620984483

Götz, F. M., Maertens, R., & van der Linden, S. (2021). *Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development*. PsyArXiv. https://doi.org/10.31234/osf.io/m6s28

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. https://doi.org/10.1037/a0021847

Graves, L., & Cherubini, F. (2016). *The rise of fact-checking sites in Europe*. Reuters Institute for the Study of Journalism. https://ora.ox.ac.uk/objects/uuid:d55ef650-e351-4526-b942-6c9e00129ad7

Greifeneder, R., Jaffé, M., Newman, E., & Schwarz, N. (2020). *The psychology of fake news: Accepting, sharing, and correcting misinformation*. Routledge. https://www.routledge.com/p/book/9780367271831

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guadagnoli, E., & Velicer, W. F. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265–275. https://doi.org/10.1037/0033-2909.103.2.265

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Guess, A., & Coppock, A. (2018). Does counter-attitudinal information cause backlash? Results from three large survey experiments. *British Journal of Political Science*, 1–19. https://doi.org/10.1017/S0007123418000327

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

Guillou, P. (2020, July 14). *Faster than training from scratch — Fine-tuning the English GPT-2 in any language with Hugging Face and fastai v2 (practical case with Portuguese)*. Medium. https://medium.com/@pierre_guillou/faster-than-training-from-scratch-fine-tuning-the-english-gpt-2-in-any-language-with-hugging-f2ec05c98787

Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Pearson. https://www.pearson.com/uk/educators/higher-education-educators/program/Hair-Multivariate-Data-Analysis-Global-Edition-7th-Edition/PGM916641.html

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. P. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116. https://doi.org/10.1037/a0038889

Hardt, O., Nader, K., & Nadel, L. (2013). Decay happens: The role of active forgetting in memory. *Trends in Cognitive Sciences*, *17*(3), 111–120. https://doi.org/10.1016/j.tics.2013.01.001

Hart, P. S., & Nisbet, E. C. (2012). Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies. *Communication Research*, *39*(6), 701–723. https://doi.org/10.1177/0093650211416646

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hass, R. G., & Grady, K. (1975). Temporal delay, type of forewarning, and resistance to influence. *Journal of Experimental Social Psychology*, *11*(5), 459–469. https://doi.org/10.1016/0022-1031(75)90048-7

Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. *Cognitive Research: Principles and Implications*, *6*, Article 38. https://doi.org/10.1186/s41235-021-00301-5

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment*, *7*(3), 238–247. https://doi.org/10.1037/1040-3590.7.3.238

Heinsohn, T., Fatke, M., Israel, J., Marschall, S., & Schultze, M. (2019). Effects of voting advice applications during election campaigns: Evidence from a panel study at the 2014 European elections. *Journal of Information Technology & Politics*, *16*(3), 250–264. https://doi.org/10.1080/19331681.2019.1644265

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Beyond WEIRD: Towards a broad-based behavioral science. *The Behavioral and Brain Sciences*, *33*(2-3), 111–135. https://doi.org/10.1017/S0140525X10000725

Hill, S. J., Lo, J., Vavreck, L., & Zaller, J. (2013). How quickly we forget: The duration of persuasion effects from mass communication. *Political Communication*, *30*(4), 521–547. https://doi.org/10.1080/10584609.2013.828143

Ho, A. K., Sidanius, J., Kteily, N., Sheehy-Skeffington, J., Pratto, F., Henkel, K. E., Foels, R., & Stewart, A. L. (2015). The nature of social dominance orientation: Theorizing and measuring preferences for intergroup inequality using the new SDO_7 scale. *Journal of Personality and Social Psychology*, *109*(6), 1003–1028. https://doi.org/10.1037/pspi0000033

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. https://psycnet.apa.org/record/1993-97193-000

Hornsey, M. J., & Fielding, K. S. (2017). Attitude roots and Jiu Jitsu persuasion: Understanding and overcoming the motivated rejection of science. *The American Psychologist*, *72*(5), 459–473. https://doi.org/10.1037/a0040437

Hotez, P., Batista, C., Ergonul, O., Figueroa, J. P., Gilbert, S., Gursel, M., Hassanain, M., Kang, G., Kim, J. H., Lall, B., Larson, H., Naniche, D., Sheahan, T., Shoham, S., Wilder-Smith, A., Strub-Wourgaft, N., Yadav, P., & Bottazzi, M. E. (2021). Correcting COVID-19 vaccine misinformation. *EClinicalMedicine*, *33*, Article 100780. https://doi.org/10.1016/j.eclinm.2021.100780

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Huck, S. W., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: a potentially confusing task. *Psychological Bulletin*, *82*(4), 511. https://doi.org/10.1037/h0076767

Insko, C. A. (1967). *Theories of attitude change*. Appleton-Century-Crofts.

Ivanov, B. (2012). Designing inoculation messages for health communication campaigns. In H. Cho (Ed.), *Health communication message design: Theory and practice* (pp. 73–93). Sage.

Ivanov, B. (2017). Inoculation theory applied in health and risk messaging. In *The Oxford Encyclopedia of Health and Risk Message Design and Processing*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.254

Ivanov, B., Miller, C. H., Compton, J., Averbeck, J. M., Harrison, K. J., Sims, J. D., Parker, K. A., & Parker, J. L. (2012). Effects of postinoculation talk on resistance to influence. *The Journal of Communication*, *62*(4), 701–718. https://doi.org/10.1111/j.1460-2466.2012.01658.x

Ivanov, B., Parker, K. A., & Dillingham, L. L. (2018). Testing the limits of inoculation-generated resistance. *Western Journal of Speech Communication*, *82*(5), 648–665. https://doi.org/10.1080/10570314.2018.1454600

Ivanov, B., & Parrott, R. (2017). Inoculation theory applied in health and risk messaging. In *The Oxford Encyclopedia of Health and Risk Message Design and Processing*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.254

Ivanov, B., Pfau, M., & Parker, K. A. (2009). Can inoculation withstand multiple attacks?: An examination of the effectiveness of the inoculation strategy compared to the supportive and restoration strategies. *Communication Research*, *36*(5), 655–676. https://doi.org/10.1177/0093650209338909

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science*, *59*(3), 690–707. https://doi.org/10.1111/ajps.12152

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

Jolley, D., & Paterson, J. L. (2020). Pylons ablaze: Examining the role of 5G COVID-19 conspiracy beliefs and support for violence. *British Journal of Social Psychology*, *59*(3), 628–640. https://doi.org/10.1111/bjso.12394

Jost, J. T., van der Linden, S., Panagopoulos, C., & Hardin, C. D. (2018). Ideological asymmetries in conformity, desire for shared reality, and the spread of misinformation. *Current Opinion in Psychology*, *23*, 77–83. https://doi.org/10.1016/j.copsyc.2018.01.003

Jung, M. K., Jeong, S. D., Noh, J. Y., Kim, D. U., Jung, S., Song, J. Y., ... & Shin, E. C. (2022). BNT162b2-induced memory T cells respond to the Omicron variant with preserved polyfunctionality. *Nature Microbiology*, *7*(6), 909–917. https://doi.org/10.1038/s41564-022-01123-x

Kahan, D. M., Jenkins-Smith, H., & Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research*, *14*(2), 147–174. https://doi.org/10.1080/13669877.2010.511246

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, *1*(1), 54–86. https://doi.org/10.1017/bpp.2016.2

Karpicke, J. D., & Roediger, H. L. III. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kerr, J. R., & Wilson, M. S. (2018). Changes in perceived scientific consensus shift beliefs about climate change and GM food safety. *PloS One*, *13*(7), e0200295. https://doi.org/10.1371/journal.pone.0200295

Koehler, D. J. (2016). Can journalistic "false balance" distort public perception of consensus in expert opinion? *Journal of Experimental Psychology: Applied*, *22*(1), 24–38. https://doi.org/10.1037/xap0000073

Konrath, S., Meier, B. P., & Bushman, B. J. (2014). Development and validation of the Single Item Narcissism Scale (SINS). *PloS One*, *9*(8), Article e103469. https://doi.org/10.1371/journal.pone.0103469

Kortenkamp, K. V., & Basten, B. (2015). Environmental science in the media: Effects of opposing viewpoints on risk and uncertainty perceptions. *Science Communication*, *37*(3), 287–313. https://doi.org/10.1177/1075547015574016

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Laato, S., Islam, A. N., Islam, M. N., & Whelan, E. (2020). What drives unverified information sharing and cyberchondria during the COVID-19 pandemic? *European Journal of Information Systems*, 1-18. https://doi.org/10.1080/0960085X.2020.1770632

Landrum, A. R., & Olshansky, A. (2019). The role of conspiracy mentality in denial of science and susceptibility to viral deception about science. *Politics and the Life*

Sciences: The Journal of the Association for Politics and the Life Sciences, *38*(2), 193–209. https://doi.org/10.1017/pls.2019.9

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

Lecheler, S., & de Vreese, C. H. (2016). How long do news framing effects last? A systematic review of longitudinal studies. *Annals of the International Communication Association*, *40*(1), 3–30. https://doi.org/10.1080/23808985.2015.11735254

Levitt, H. M., Bamberg, M., Creswell, J. W., Frost, D. M., Josselson, R., & Suárez-Orozco, C. (2018). Journal article reporting standards for qualitative primary, qualitative meta-analytic, and mixed methods research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 26–46. https://doi.org/10.1037/amp0000151

Lewandowsky, S. (2020). The 'post-truth' world, misinformation, and information literacy: A perspective from cognitive science. In S. Goldstein (Ed.), *Informed Societies* (pp. 69–88). Facet Publishing. https://doi.org/10.29085/9781783303922.006

Lewandowsky, S., Cook, J., & Ecker, U. K. H. (2017). Letting the gorilla emerge from the mist: Getting past post-truth. *Journal of Applied Research in Memory and Cognition*, *6*(4), 418–424. https://doi.org/10.1016/j.jarmac.2017.11.002

Lewandowsky, S., Cook, J., Ecker, U. K. H., Albarracín, D., Amazeen, M. A., Kendeou, P., Lombardi, D., Newman, E. J., Pennycook, G., Porter, E. Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G. M.,

Swire-Thompson, B., van der Linden, S., Vraga, E. K., Wood, T. J., Zaragoza, M. S. (2020). The Debunking Handbook 2020. https://doi.org/10.17910/b7.1182

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the 'post-truth' era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

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: A Journal of the American Psychological Society*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Gignac, G. E., & Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change*, *3*(4), 399–404. https://doi.org/10.1038/nclimate1720

Lewandowsky, S., & Li, S.-C. (1995). Catastrophic interference in neural networks: Causes, solutions, and data. In F. N. Dempster, C. J. Brainerd, & C. J. Brainerd (Eds.), *Interference and Inhibition in Cognition* (pp. 329–361). Academic Press. https://doi.org/10.1016/B978-012208930-5/50011-8

Lewandowsky, S., Oreskes, N., Risbey, J. S., Newell, B. R., & Smithson, M. (2015). Seepage: Climate change denial and its effect on the scientific community. *Global Environmental Change: Human and Policy Dimensions*, *33*, 1–13. https://doi.org/10.1016/j.gloenvcha.2015.02.013

Lewandowsky, S., Pilditch, T. D., Madsen, J. K., Oreskes, N., & Risbey, J. S. (2019). Influence and seepage: An evidence-resistant minority can affect public opinion and scientific belief formation. *Cognition*, *188*, 124–139. https://doi.org/10.1016/j.cognition.2019.01.011

Lewandowsky, S., Smillie, L., Garcia, D., Hertwig, R., Weatherall, J., Egidy, S., Robertson, R. E., O'Connor, C., Kozyreva, A., Lorenz-Spreen, P., Blaschke, Y., & Leiser, M. R. (2020). *Technology and democracy: Understanding the influence of online technologies on political behaviour and decision-making*. Publications Office of the European Union. https://doi.org/10.2760/709177

Lewandowsky, S., Stritzke, W. G. K., Freund, A. M., Oberauer, K., & Krueger, J. I. (2013). Misinformation, disinformation, and violent conflict: From Iraq and the 'War on Terror' to future threats to peace. *The American Psychologist*, *68*(7), 487–501. https://doi.org/10.1037/a0034515

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lim, J. S., & Ki, E. J. (2007). Resistance to ethically suspicious parody video on YouTube: A test of inoculation theory. *Journalism & Mass Communication Quarterly*, *84*(4), 713–728. https://journals.sagepub.com/doi/abs/10.1177/107769900708400404

Linton, M. (1975). Memory for real-world events. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition*. Freeman.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and

USA. *Nature Human Behaviour*, *5*(3), 337–348. https://doi.org/10.1038/s41562-021-01056-1

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Maertens, R., Anseel, F., & van der Linden, S. (2020). Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects. *Journal of Environmental Psychology*, *70*, Article 101455. https://doi.org/10.1016/j.jenvp.2020.101455

Maertens, R., Götz, F. M., Schneider, C. R., Roozenbeek, J., Kerr, J. R., Stieger, S., McClanahan, W. P., Drabot, K., & van der Linden, S. (2022). *The Misinformation Susceptibility Test (MIST): A psychometrically validated measure of news veracity discernment*. PsyArXiv. https://doi.org/10.31234/osf.io/gk68h

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1–16. https://doi.org/10.1037/xap0000315

Markon, K. E. (2019). Bifactor and hierarchical models: Specification, inference, and interpretation. *Annual Review of Clinical Psychology*, *15*, 51–69. https://doi.org/10.1146/annurev-clinpsy-050718-095522

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*, *5*(1), 1–20. https://doi.org/10.1186/s41235-020-00252-3

McCright, A. M., Charters, M., Dentzman, K., & Dietz, T. (2016). Examining the effectiveness of climate change frames in the face of a climate change denial counter-frame. *Topics in Cognitive Science*, *8*(1), 76–97. https://doi.org/10.1111/tops.12171

McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press. https://doi.org/10.4324/9781410601087

McGuire, W. J. (1961a). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *Journal of Abnormal and Social Psychology*, *63*(2), 326–332. https://doi.org/10.1037/h0048344

McGuire, W. J. (1961b). The effectiveness of supportive and refutational defenses in immunizing and restoring beliefs against persuasion. *Sociometry*, *24*(2), 184–197. https://www.jstor.org/stable/pdf/2786067.pdf

McGuire, W. J. (1962). Persistence of the resistance to persuasion induced by various types of prior belief defenses. *The Journal of Abnormal and Social Psychology*, *64*(4), 241–248. https://doi.org/10.1037/h0044167

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. *Advances in Experimental Social Psychology*, *1*, 191–229. https://doi.org/10.1016/S0065-2601(08)60052-0

McGuire, W. J. (1970). Vaccine for brainwash. *Psychology Today*, *3*(9), 36–39.

McGuire, W. J. (1973). Persuasion, resistance, and attitude change. In I. de Sola Pool, W. Schramm, F. W. Frey, & E. B. Parker (Eds.), *Handbook of communication* (pp. 216–252). Rand McNally College Publishing Company.

McGuire, W. J., & Papageorgis, D. (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *Journal of Abnormal and Social Psychology*, *62*(2), 327–337. https://doi.org/10.1037/h0042026

McGuire, W. J., & Papageorgis, D. (1962). Effectiveness of forewarning in developing resistance to persuasion. *Public Opinion Quarterly*, *26*(1), 24–34. https://doi.org/10.1086/267068

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Michel, N., Cater, J. J., III, & Varela, O. (2009). Active versus passive teaching styles: An empirical study of student learning outcomes. *Human Resource Development Quarterly*, *20*(4), 397–418. https://doi.org/10.1002/hrdq.20025

Miller, C. H., Ivanov, B., Sims, J., Compton, J., Harrison, K. J., Parker, K. A., Parker, J. L., & Averbeck, J. M. (2013). Boosting the potency of resistance: Combining the motivational forces of inoculation and psychological reactance. *Human Communication Research*, *39*(1), 127–155. https://doi.org/10.1111/j.1468-2958.2012.01438.x

Motyl, M., Iyer, R., Oishi, S., Trawalter, S., & Nosek, B. A. (2014). How ideological migration geographically segregates groups. *Journal of Experimental Social Psychology*, *51*, 1–14. https://doi.org/10.1016/j.jesp.2013.10.010

Murre, J. M. J., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PloS One*, *10*(7), e0120644. https://doi.org/10.1371/journal.pone.0120644

Murty, V. P., & Dickerson, K. C. (2016). Motivational influences on memory. In S. I. Kim, J. Reeve, & M. Bong (Eds.), *Recent developments in neuroscience research on human motivation* (pp. 203–227). Emerald Group Publishing. https://doi.org/10.1108/S0749-742320160000019019

Nabi, R. L. (2003). " Feeling" resistance: Exploring the role of emotionally evocative visuals in inducing inoculation. *Media Psychology*, *5*(2), 199–223. https://doi.org/10.1207/S1532785XMEP0502_4

Nader, K., & Hardt, O. (2009). A single standard for memory: the case for reconsolidation. *Nature Reviews Neuroscience*, *10*(3), 224–234. https://doi.org/10.1038/nrn2590

Nasser, M. A. (2020, March 26). *Step-by-step guide on how to train GPT-2 on books using Google Colab*. Towards Data Science. https://towardsdatascience.com/step-by-step-guide-on-how-to-train-gpt-2-on-books-using-google-colab-b3c6fa15fef0

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society*, *20*(10), 3720–3737.

Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The Patient*, *7*(1), 23–35. https://doi.org/10.1007/s40271-013-0041-0

Niederdeppe, J., Heley, K., & Barry, C. L. (2015). Inoculation and narrative strategies in competitive framing of three health policy issues. *The Journal of Communication*, *65*(5), 838–862. https://doi.org/10.1111/jcom.12162

Nisa, C. F., Bélanger, J. J., Schumpe, B. M., & Faller, D. G. (2019). Meta-analysis of randomised controlled trials testing behavioural interventions to promote household action on climate change. *Nature Communications*, *10*, Article 4545. https://doi.org/10.1038/s41467-019-12457-2

Norenzayan, A., & Hansen, I. G. (2006). Belief in supernatural agents in the face of death. *Personality & Social Psychology Bulletin*, *32*(2), 174–187. https://doi.org/10.1177/0146167205280251

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nugent, C. (2018). How to stop WhatsApp's deadly fake news crisis in india. *Time*. https://time.com/5352516/india-whatsapp-fake-news/

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2020). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*, *42*(3), 939–960. https://doi.org/10.1007/s11109-019-09528-x

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review*, *115*(3), 544–576. https://doi.org/10.1037/0033-295X.115.3.544

Oreskes, N. (2004). The scientific consensus on climate change. *Science*, *306*(5702), 1686–1686. https://doi.org/10.1126/science.1103618

Oreskes, N., & Conway, E. M. (2010). Defeating the merchants of doubt. *Nature*, *465*(7299), 686–687. https://doi.org/10.1038/465686a

Oreskes, N., & Conway, E. M. (2011). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. Bloomsbury Publishing USA.

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Papageorgis, D., & McGuire, W. J. (1961). The generality of immunity to persuasion produced by pre-exposure to weakened counterarguments. *Journal of Abnormal and Social Psychology*, *62*, 475–481. https://doi.org/10.1037/h0048430

Parker, K. A., Rains, S. A., & Ivanov, B. (2016). Examining the 'blanket of protection' conferred by inoculation: The effects of inoculation messages on the cross-protection of related attitudes. *Communication Monographs*, *83*(1), 49–68. https://doi.org/10.1080/03637751.2015.1030681

Paulhus, D. L., Buckels, E. E., Trapnell, P. D., & Jones, D. N. (2020). Screening for dark personalities. *European Journal of Psychological Assessment*, *37*(3), 208–222. https://doi.org/10.1027/1015-5759/a000602

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. https://doi.org/10.1016/j.jesp.2017.01.006

Pennycook, G., Binnendyk, J., Newton, C., & Rand, D. G. (2021). A practical guide to doing behavioral research on fake news and misinformation. *Collabra: Psychology*, *7*(1), Article 25293. https://doi.org/10.1525/collabra.25293

Pennycook, G., Cannon, T., & 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., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, *10*(6), 549–563. http://journal.sjdm.org/15/15923a/jdm15923a.pdf

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

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

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

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*. https://doi.org/10.1111/jopy.12476

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Petersen, A. M., Vincent, E. M., & Westerling, A. L. (2019). Discrepancy in scientific authority and media visibility of climate change scientists and contrarians. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09959-4

Petty, R. E., & Cacioppo, J. T. (1979). Issue involvement can increase or decrease persuasion by enhancing message-relevant cognitive responses. *Journal of Personality and Social Psychology*, *37*(10), 1915–1926. https://doi.org/10.1037/0022-3514.37.10.1915

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of persuasion. In R. E. Petty & J. T. Cacioppo (Eds.), *Communication and persuasion: Central and peripheral routes to attitude change* (pp. 1–24). Springer. https://doi.org/10.1007/978-1-4612-4964-1_1

Petty, R. E., Priester, J. R., & Wegener, D. T. (1994). Cognitive processes in attitude change. In R. S. Wyer Jr. & T. K. Srull (Eds.), *Handbook of social cognition: Basic processes; Applications* (pp. 69–142). Lawrence Erlbaum Associates, Inc.

Pew Research Center (2020, September 28). *Many americans get news on YouTube, where news organizations and independent producers thrive side by side* [Report]. https://www.pewresearch.org/journalism/2020/09/28/many-americans-get-news-on-youtube-where-news-organizations-and-independent-producers-thrive-side-by-side/

Pew Research Center (2021, April 7). *Social media use in 2021* [Report]. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Pfau, M. (1992). The potential of inoculation in promoting resistance to the effectiveness of comparative advertising messages. *Communication Quarterly*, *40*(1), 26–44. https://doi.org/10.1080/01463379209369818

Pfau, M., Compton, J., Parker, K. A., An, C., Wittenberg, E. M., Ferguson, M., Horton, H., & Malyshev, Y. (2006). The conundrum of the timing of counterarguing effects in resistance: Strategies to boost the persistence of counterarguing output. *Communication Quarterly*, *54*(2), 143–156. https://doi.org/10.1080/01463370600650845

Pfau, M., Compton, J., Parker, K. A., Wittenberg, E. M., An, C., Ferguson, M., Horton, H., & Malyshev, Y. (2004). The traditional explanation for resistance versus attitude accessibility: Do they trigger distinct or overlapping processes of resistance? *Human Communication Research*, *30*(3), 329–360. https://doi.org/10.1111/j.1468-2958.2004.tb00735.x

Pfau, M., Holbert, R. L., Zubric, S. J., Pasha, N. H., & Lin, W. K. (2000). Role and influence of communication modality in the process of resistance to persuasion. *Media Psychology*, *2*(1), 1–33. https://doi.org/10.1207/S1532785XMEP0201_1

Pfau, M., Ivanov, B., Houston, B., Haigh, M., Sims, J., Gilchrist, E., Russell, J., Wigley, S., Eckstein, J., & Richert, N. (2005). Inoculation and mental processing: The instrumental role of associative networks in the process of resistance to counterattitudinal influence. *Communication Monographs*, *72*(4), 414–441. https://doi.org/10.1080/03637750500322578

Pfau, M., Roskos-Ewoldsen, D., Wood, M., Yin, S., Cho, J., Lu, K.-H., & Shen, L. (2003). Attitude accessibility as an alternative explanation for how inoculation confers resistance. *Communication Monographs*, *70*(1), 39–51. https://doi.org/10.1080/715114663

Pfau, M., Szabo, A., Anderson, J., Morrill, J., Zubric, J., & Wan, H. H. (2001). The role and impact of affect in the process of resistance to persuasion. *Human Communication Research*, *27*(2), 216–252. https://doi.org/10.1111/j.1468-2958.2001.tb00781.x

Pfau, M., Tusing, K. J., Koerner, A. F., Lee, W., Godbold, L. C., Penaloza, L. J., Yang, V. S.-H., & Hong, Y.-H. (1997). Enriching the inoculation construct: The role of critical components in the process of resistance. *Human Communication Research*, *24*(2), 187–215. https://doi.org/10.1111/j.1468-2958.1997.tb00413.x

Pfau, M., & Van Bockern, S. (1994). The persistence of inoculation in conferring resistance to smoking initiation among adolescents. *Human Communication Research*, *20*(3), 413–430. https://doi.org/10.1111/j.1468-2958.1994.tb00329.x

Pfau, M., Van Bockern, S., & Kang, J. G. (1992). Use of inoculation to promote resistance to smoking initiation among adolescents. *Communication Monographs*, *59*(3), 213–230. https://doi.org/10.1080/03637759209376266

Piazza, J. A. (2022). Fake news: The effects of social media disinformation on domestic terrorism. *Dynamics of Asymmetric Conflict*, *15*(1), 55–77. https://doi.org/10.1080/17467586.2021.1895263

Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*. Routledge. https://doi.org/10.4324/9781315814919

Pryor, B., & Steinfatt, T. M. (1978). The effects of initial belief level on inoculation theory and its proposed mechanisms. *Human Communication Research*, *4*(3), 217–230. https://doi.org/10.1111/j.1468-2958.1978.tb00611.x

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

Rammstedt, B., Lechner, C. M., & Danner, D. (2021). Short forms do not fall short: A comparison of three (extra-)short forms of the Big Five. *European Journal of Psychological Assessment*, *37*(1), 23–32. https://doi.org/10.1027/1015-5759/a000574

Readfearn, G. (2016). *Revealed: Most popular climate story on social media told half a million people the science was a hoax*. https://www.desmogblog.com/2016/11/29/revealed-most-popular-climate-story-social-media-told-half-million-people-science-was-hoax

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552–566. https://doi.org/10.1037/0033-2909.114.3.552

Rentfrow, P. J., Gosling, S. D., Jokela, M., Stillwell, D. J., Kosinski, M., & Potter, J. (2013). Divided we stand: Three psychological regions of the United States and their political, economic, social, and health correlates. *Journal of Personality and Social Psychology*, *105*(6), 996–1012. https://doi.org/10.1037/a0034434

Rentfrow, P. J., Jokela, M., & Lamb, M. E. (2015). Regional personality differences in Great Britain. *PloS One*, *10*(3), Article e0122245. https://doi.org/10.1371/journal.pone.0122245

Revelle, W. (2021). psych: Procedures for psychological, psychometric, and personality research. *The Comprehensive R Archive Network*. https://cran.r-project.org/package=psych

Revelle, W., & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological Assessment*, *31*(12), 1395–1411. https://doi.org/10.1037/pas0000754

Richards, A. S., & Banas, J. A. (2018). The opposing mediational effects of apprehensive threat and motivational threat when inoculating against reactance to health promotion. *Southern Communication Journal*, *83*(4), 245–255. https://doi.org/10.1080/1041794X.2018.1498909

Richards, A. S., Banas, J. A., & Magid, Y. (2017). More on inoculating against reactance to persuasive health messages: The paradox of threat. *Health Communication*, *32*(7), 890–902. https://doi.org/10.1080/10410236.2016.1196410

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg self-esteem scale. *Personality & Social Psychology Bulletin*, *27*(2), 151–161. https://doi.org/10.1177/0146167201272002

Roediger, H. L. III, & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roediger, H. L. III, & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Rogers, R. W., & Thistlethwaite, D. L. (1969). An analysis of active and passive defenses in inducing resistance to persuasion. *Journal of Personality and Social Psychology*, *11*(4), 301–308. https://doi.org/10.1037/h0027354

Roozenbeek, J., Freeman, A. L. J., & van der Linden, S. (2021). How accurate are accuracy-nudge interventions? A preregistered direct replication of Pennycook et al. (2020). *Psychological Science*, *32*(7), 1169–1178. https://doi.org/10.1177/09567976211024535

Roozenbeek, J., Maertens, R., Herzog, S. M., Geers, M., Kurvers, R. H., Sultan, M., & van der Linden, S. (2022). Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, *17*(3), 547–573. https://journal.sjdm.org/22/220228/jdm220228.html

Roozenbeek, J., Maertens, R., McClanahan, W., & van der Linden, S. (2021). 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., Schneider, C. R., Dryhurst, S., Kerr, J., Freeman, A. L. J., Recchia, G., van der Bles, A. M., & van der Linden, S. (2020). Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science*, *7*(10), Article 201199. https://doi.org/10.1098/rsos.201199

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

Roozenbeek, J., & van der Linden, S. (2019a). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), Article 65. https://doi.org/10.1057/s41599-019-0279-9

Roozenbeek, J., & van der Linden, S. (2019b). The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570–580. https://doi.org/10.1080/13669877.2018.1443491

Roozenbeek, J., & van der Linden, S. (2020). Breaking Harmony Square: A game that "inoculates" against political misinformation. *Harvard Kennedy School Misinformation Review*, *1*(8), 1–26. https://doi.org/10.37016/mr-2020-47

Roozenbeek, J., van der Linden, S., Goldberg, B., Rathje, S., & Lewandowsky, S. (2022). Psychological inoculation improves resilience against misinformation on social media. *Science Advances*, *8*(34), Article eabo6254. https://doi.org/10.1126/sciadv.abo6254

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on the psychological theory of 'inoculation' can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016//mr-2020-008

Rosellini, A. J., & Brown, T. A. (2021). Developing and validating clinical questionnaires. *Annual Review of Clinical Psychology*, *17*, 55–81. https://doi.org/10.1146/annurev-clinpsy-081219-115343

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sanderson, J. A., Gignac, G. E., & Ecker, U. K. (2021). Working memory capacity, removal efficiency and event specific memory as predictors of misinformation reliance.

Journal of Cognitive Psychology, *33*(5), 518–532. https://doi.org/10.1080/20445911.2021.1931243

Schaffner, B. F., & Luks, S. (2018). Misinformation or expressive responding? What an inauguration crowd can tell us about the source of political misinformation in surveys. *Public Opinion Quarterly*, *82*(1), 135–147. https://doi.org/10.1093/poq/nfx042

Scheufele, D. A., & Krause, N. M. (2019). Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1805871115

Schumacker, R. E., Lomax, R. G., & Schumacker, R. (2015). *A beginner's guide to structural equation modeling* (4th ed.). Routledge. https://www.routledge.com/A-Beginners-Guide-to-Structural-Equation-Modeling-Fourth-Edition/Schumacker-Lomax-Schumacker-Lomax/p/book/9781138811935

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*(11), 966–972. https://doi.org/10.7326/0003-4819-127-11-199712010-00003

Simms, L. J. (2008). Classical and modern methods of psychological scale construction. *Social and Personality Psychology Compass*, *2*(1), 414–433. https://doi.org/10.1111/j.1751-9004.2007.00044.x

Smith, E. R. (1998). Mental representation and memory. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 391–445). McGraw-Hill.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin*, *46*(2), 137–150. https://doi.org/10.1037/h0062958

Song, M.-K., & Ward, S. E. (2015). Assessment effects in educational and psychosocial intervention trials: An important but often-overlooked problem. *Research in Nursing & Health*, *38*(3), 241–247. https://doi.org/10.1002/nur.21651

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality*, *68*, 69–81. https://doi.org/10.1016/j.jrp.2017.02.004

Steiner, M., & Grieder, S. (2020). EFAtools: An R package with fast and flexible implementations of exploratory factor analysis tools. *Journal of Open Source Software*, *5*(53), Article 2521. https://doi.org/10.21105/joss.02521

Stenhouse, N., Maibach, E., Cobb, S., Ban, R., Bleistein, A., Croft, P., Bierly, E., Seitter, K., Rasmussen, G., & Leiserowitz, A. A. (2014). Meteorologists' views about global warming: A survey of american meteorological society professional members. *Bulletin of the American Meteorological Society*, *95*(7), 1029–1040. https://doi.org/10.1175/bams-d-13-00091.1

Stenhouse, N., Myers, T. A., Vraga, E. K., Kotcher, J. E., Beall, L., & Maibach, E. W. (2018). The potential role of actively open-minded thinking in preventing motivated reasoning about controversial science. *Journal of Environmental Psychology*, *57*, 17–24. https://doi.org/10.1016/j.jenvp.2018.06.001

Stocker, T. F., Qin, D., Plattner, G. K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., & Midgley, P. M. (2013). *Climate change 2013: The physical science basis*. IPCC. https://www.ipcc.ch/site/assets/uploads/2018/02/WG1AR5_all_final.pdf

Stocking, S. H., Holly Stocking, S., & Holstein, L. W. (2009). Manufacturing doubt: Journalists' roles and the construction of ignorance in a scientific controversy. *Public Understanding of Science*, *18*(1), 23–42. https://doi.org/10.1177/0963662507079373

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Swami, V., Chamorro-Premuzic, T., & Furnham, A. (2010). Unanswered questions: A preliminary investigation of personality and individual difference predictors of 9/11 conspiracist beliefs. *Applied Cognitive Psychology*, *24*(6), 749–761. https://doi.org/10.1002/acp.1583

Swire, B., Berinsky, A. J., Lewandowsky, S., & Ecker, U. K. H. (2017). Processing political misinformation: Comprehending the Trump phenomenon. *Royal Society Open Science*, *4*(3), Article 160802. https://doi.org/10.1098/rsos.160802

Swire, B., & Ecker, U. K. H. (2018). Misinformation and its correction: Cognitive mechanisms and recommendations for mass communication. In B. G. Southwell, E. A. Thorson, & L. Sheble (Eds.), *Misinformation and mass audiences* (pp. 195–211). University of Texas Press. https://doi.org/10.7560/314555-013

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *43*(12), 1948–1961. https://doi.org/10.1037/xlm0000422

Swire-Thompson, B., DeGutis, J., & Lazer, D. (2020). Searching for the backfire effect: Measurement and design considerations. *Journal of Applied Research in Memory and Cognition*, *9*(3), 286–299. https://doi.org/10.1016/j.jarmac.2020.06.006

Swire-Thompson, B., Miklaucic, N., Wihbey, J. P., Lazer, D., & DeGutis, J. (2022). The backfire effect after correcting misinformation is strongly associated with reliability. *Journal of Experimental Psychology: General*, *151*(7), 1655–1665. https://doi.org/10.1037/xge0001131

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Pearson. https://psycnet.apa.org/record/2006-03883-000

Tam, K.-P., & Milfont, T. L. (in press). Cross-cultural environmental psychology. *Journal of Experimental Psychology*. https://osf.io/tva7f

Tannenbaum, P. H., Macauley, J. R., & Norris, E. L. (1966). Principle of congruity and reduction of persuasion. *Journal of Personality and Social Psychology*, *3*(2), 233–238. https://doi.org/10.1037/h0022893

Tay, L. Q., Hurlstone, M. J., Kurz, T., & Ecker, U. K. H. (2022). 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

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief to medium-length Big Five and Big Six Personality Questionnaires. *Psychological Assessment*, *23*(4), 995–1009. https://doi.org/10.1037/a0024165

Theußl, S., Feinerer, I., & Hornik, K. (2012). A tm plug-in for distributed text mining in R. *Journal of Statistical Software*, *51*, 1–31. https://doi.org/10.18637/jss.v051.i05

Thorndike, E. L. (1913). *The psychology of learning*. Teachers College, Columbia University.

Thurstone, L. L. (1944). Second-order factors. *Psychometrika*, *9*(2), 71–100. https://doi.org/10.1007/BF02288715

Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, *700*(1), 136–151. https://doi.org/10.1177/00027162221087936

Tyng, C. M., Amin, H. U., Saad, M. N., & Malik, A. S. (2017). The influences of emotion on learning and memory. *Frontiers in Psychology*, *8*, Article 1454. https://doi.org/10.3389/fpsyg.2017.01454

Uenal, F., Sidanius, J., Maertens, R., Hudson, S. T. J., Davis, G., & Ghani, A. (2022). The roots of ecological dominance orientation: Assessing individual preferences for an anthropocentric and hierarchically organized world. *Journal of Environmental Psychology*, *81*, Article 101783. https://doi.org/10.1016/j.jenvp.2022.101783

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*(1), 49–60. https://doi.org/10.1037/h0044616

United States Census Bureau (2019). *Age and sex composition in the United States* [Data set]. https://www.census.gov/data/tables/2019/demo/age-and-sex/2019-age-sex-compositio n.html

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review*, *15*(1), 84–113. https://doi.org/10.1111/sipr.12077

van der Linden, S. (2019). Countering science denial. *Nature Human Behaviour*, *3*, 889–890. https://doi.org/10.1038/s41562-019-0631-5

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, *28*(3), 460–467. https://doi.org/10.1038/s41591-022-01713-6

van der Linden, S., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2014). How to communicate the scientific consensus on climate change: Plain facts, pie charts or metaphors? *Climatic Change*, *126*(1), 255–262. https://doi.org/10.1007/s10584-014-1190-4

van der Linden, S., Leiserowitz, A. A., Feinberg, G. D., & Maibach, E. W. (2015). The scientific consensus on climate change as a gateway belief: Experimental evidence. *PloS One*, *10*(2), Article e0118489. https://doi.org/10.1371/journal.pone.0118489

van der Linden, S., Leiserowitz, A. A., & Maibach, E. (2018). Scientific agreement can neutralize politicization of facts. *Nature Human Behaviour*, *2*(1), 2–3. https://doi.org/10.1038/s41562-017-0259-2

van der Linden, S., Leiserowitz, A. A., & Maibach, E. (2019). The gateway belief model: A large-scale replication. *Journal of Environmental Psychology*, *62*, 49–58. https://doi.org/10.1016/j.jenvp.2019.01.009

van der Linden, S., Leiserowitz, A. A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges*, *1*(2), Article 1600008. https://doi.org/10.1002/gch2.201600008

van der Linden, S., Maibach, E., & Leiserowitz, A. A. (2015). Improving public engagement with climate change: Five "best practice" insights from psychological science. *Perspectives on Psychological Science*, *10*(6), 758–763. https://doi.org/10.1177/1745691615598516

van der Linden, S., Maibach, E., & Leiserowitz, A. A. (2019). Exposure to scientific consensus does not cause psychological reactance. *Environmental Communication*. https://doi.org/10.1080/17524032.2019.1617763

van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: political bias in perceptions of fake news. *Media Culture & Society*, *42*(3), 460–470. https://doi.org/10.1177/0163443720906992

van der Linden, S., & Roozenbeek, J. (2020). Psychological inoculation against fake news. In R. Greifeneder, M. Jaffé, E. J. Newman, & N. Schwarz (Eds.), *The psychology of fake news: Accepting, sharing, and correcting misinformation*. Routledge. https://www.routledge.com/p/book/9780367271831

van der Linden, S., Roozenbeek, J., & Compton, J. (2020). Inoculating against fake news about COVID-19. *Frontiers in Psychology*, *11*, Article 566790. https://doi.org/10.3389/fpsyg.2020.566790

van der Linden, S., Roozenbeek, J., Maertens, R., Basol, M., Kácha, O., Rathje, S., & Traberg, C. S. (2021). How can psychological science help counter the spread of fake

news? *The Spanish Journal of Psychology*, *24*, Article e25. https://doi.org/10.1017/SJP.2021.23

van Prooijen, J.-W., Krouwel, A. P. M., & Pollet, T. V. (2015). Political extremism predicts belief in conspiracy theories. *Social Psychological and Personality Science*, *6*(5), 570–578. https://doi.org/10.1177/1948550614567356

Varker, T., & Devilly, G. J. (2012). An analogue trial of inoculation/resilience training for emergency services personnel: Proof of concept. *Journal of Anxiety Disorders*, *26*(6), 696–701. https://doi.org/10.1016/j.janxdis.2012.01.009

Verheggen, B., Strengers, B., Cook, J., van Dorland, R., Vringer, K., Peters, J., Visser, H., & Meyer, L. (2014). Scientists' views about attribution of global warming. *Environmental Science & Technology*, *48*(16), 8963–8971. https://doi.org/10.1021/es501998e

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Vraga, E. K., Bode, L., & Tully, M. (2021). The effects of a news literacy video and real-time corrections to video misinformation related to sunscreen and skin cancer. *Health Communication*, 1–9. https://doi.org/10.1080/10410236.2021.1910165

Weber, R., & Popova, L. (2012). Testing equivalence in communication research: Theory and application. *Communication Methods and Measures*, *6*(3), 190–213. https://doi.org/10.1080/19312458.2012.703834

Weiner, I. B., Schinka, J. A., & Velicer, W. F. (2012). *Handbook of psychology: Research methods in psychology* (2nd ed., Vol. 2). John Wiley & Sons. https://www.wiley.com/en-gb/Handbook+of+Psychology%2C+Volume+2%2C+Research+Methods+in+Psychology%2C+2nd+Edition-p-9780470890646

Williams, M. N., & Bond, C. M. (2020). A preregistered replication of "Inoculating the public against misinformation about climate change". *Journal of Environmental Psychology*, *70*, Article 101456. https://doi.org/10.1016/j.jenvp.2020.101456

Wood, M. L. M. (2007). Rethinking the inoculation analogy: Effects on subjects with differing preexisting attitudes. *Human Communication Research*, *33*(3), 357–378. https://doi.org/10.1111/j.1468-2958.2007.00303.x

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Woolf, M. (2019, September 4). *How to make custom AI-generated text with GPT-2*. Max Woolf's Blog. https://minimaxir.com/2019/09/howto-gpt2/

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist*, *34*(6), 806–838. https://doi.org/10.1177/0011000006288127

Zerback, T., Töpfl, F., & Knöpfle, M. (2021). The disconcerting potential of online disinformation: Persuasive effects of astroturfing comments and three strategies for inoculation against them. *New Media & Society*, *23*(5), 1080–1098. https://doi.org/10.1177/1461444820908530

Zickar, M. J. (2020). Measurement development and evaluation. *Annual Review of Organizational Psychology and Organizational Behavior*, *7*, 213–232. https://doi.org/10.1146/annurev-orgpsych-012119-044957

Zihiri, S., Lima, G., Han, J., Cha, M., & Lee, W. (2022). QAnon shifts into the mainstream, remains a far-right ally. *Heliyon*, *8*(2), Article e08764. https://doi.org/10.1016/j.heliyon.2022.e08764