Effectiveness of digital health interventions against COVID-19 misinformation: a systematic realist review of intervention trials

Robert Dickinson[1¶*]

Dominique Makowski[2]

Harm van Marwijk[1]

Elizabeth Ford[1]

[1] Department of Primary Care and Public Health, Brighton and Sussex Medical School, Brighton, United Kingdom

[2] Department of Psychology, University of Sussex, Brighton, United Kingdom

Abstract

Misinformation is a growing concern worldwide, particularly in public health following the COVID-19 pandemic in which misinformation has been attributed to tens of thousands of unnecessary deaths. Therefore a search for effective interventions against misinformation is underway, with widely varying proposed interventions, measures of efficacy, and groups targeted for intervention. This realist systematic review of proposed interventions against COVID-19 misinformation assesses the studies themselves, the characteristics and effectiveness of the interventions proposed, the durability of effect, and the circumstances and contexts within which these interventions function. We searched several databases for studies testing interventions published from 2020 onwards. The search results were sorted by eligibility, with eligible studies then being coded by themes and assessed for quality. Twenty-six studies were included, representing eight types of intervention.

The results are promising to the advantages of game-type interventions, with other types scoring poorly on either scalability or impact. Backfire effects and effects on subgroups were reported on intermittently in the included studies, showing the advantages of certain interventions for subgroups or contexts. No one intervention appears sufficient by itself, therefore this study recommends the creation of packages of interventions by policymakers, who can tailor the package for contexts and targeted groups. There was high heterogeneity in outcome measures and methods, making comparisons between studies difficult; this should be a focus in future studies. Additionally, the theoretical and intervention literatures need connecting for greater understanding of the mechanisms at work in the interventions. Lastly, there is a need for work more explicitly addressing political polarisation and its role in the belief and spread of misinformation. This study contributes toward the expansion of realist review approaches, understandings of COVID-19 misinformation interventions, and broader debates around the nature of politicisation in contemporary misinformation.

Author Summary

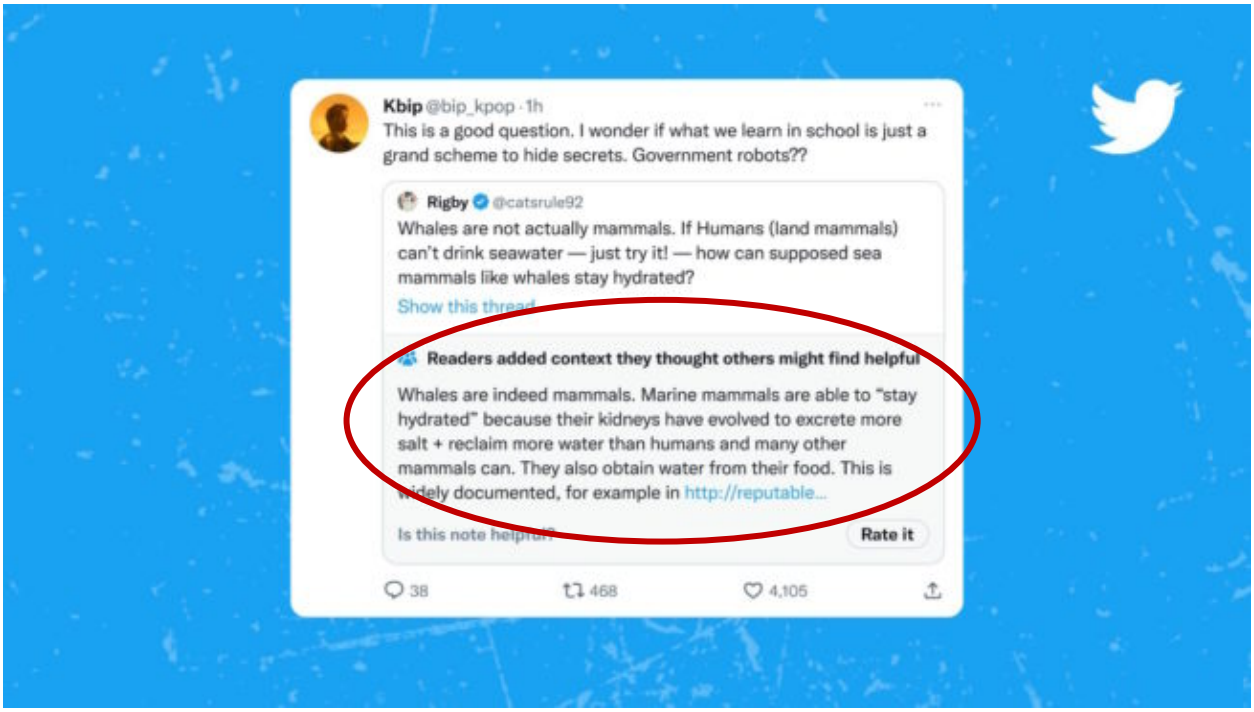
Misinformation is increasingly seen as a danger to public health and society at large. In the case of COVID-19, it is associated with high levels of unnecessary death among the public. There have been many interventions proposed to counter misinformation, yet little taking a meta-analytical perspective. These interventions vary greatly and are not measured for effect in the same ways, making traditional comparisons difficult. Instead, we categorised the interventions by type and assessed them by impact, scalability, durability, and which groups of people and contexts in which they best work. With this information for each type of intervention, policymakers can then make packages of multiple interventions that best work in their circumstances. Although game-type interventions stood out from the rest, no one intervention seems capable of effectively countering misinformation by itself. Many interventions were found to work differently on different groups of people, which reaffirms suggestions by some authors that political ideology is relevant to how people respond to these interventions. In future research there is a need to more deeply investigate the role of politicisation in misinformation and interventions against it, as well as bringing in more theory to understand how these interventions function.

Introduction

Misinformation has been a societal issue throughout history. This phenomenon can be seen in many areas, but perhaps most clearly in public health, where alongside the introduction of many major advances in medicine came movements of resistance and misinformation. In the contemporary, systemic misinformation is a well-established by-product of increasing reliance on the internet and social media for the dissemination of news and information. Public concern about misinformation appeared to reach a new height in 2016 in relation to the US Presidential Election, particularly around perceptions of misinformation campaigns supporting Donald Trump's bid for the presidency. By many accounts, this period resulted in the development of an infrastructure of misinformation, accelerated by social media algorithms to reach new and greater audiences. In 2020, when the COVID-19 pandemic began, misinformation began and continued to punctuate the public understanding of the pandemic and the public health response thereto.

The pre-COVID misinformation intervention landscape was dominated by fact-checking. Fact-checking can be described as a form of debunking in which information is retroactively checked for veracity and if found

Figure 1: X/Twitter post with crowdsourced fact-checking highlighted in red



to be inaccurate changed. Importantly, an additional step in fact-checking and other forms of debunking is attempting to reach the audience initially exposed to the misinformation and retroactively change their internalised understanding of the information [1]. Recently, accuracy nudges have been championed as a new, primary intervention-type [2]. Accuracy nudges refer to a variety of interventions that 'nudge' people to consider the veracity of the information they are seeing or are about to see. This can include prompts that appear on-screen next to links to news articles, or fact-checks that appear alongside social media posts but can also take on a wide variety of forms. Below is an example from X/Twitter highlighted by a red circle (figure 1), that shows crowdsourced fact-checking to appear next to suspected misinformation [3].

Although championed through seminal studies like [4], accuracy nudge interventions have since garnered significant criticism on their effectiveness and the potential impact of partisan bias in participants [5, 6, 7] including replication studies that did not replicate the initial findings [8].

In the theoretical literature, discussion of misinformation interventions focused on inoculation, backfire effects, and the importance of worldview in intervention effectiveness [2]. Inoculation refers to the idea of

75 priming people before they might encounter misinformation to make them more aware of it with the goal of
76 building resilience against it. The 'backfire' effect is an issue widely theorised about in the literature around
77 misinformation, typically centred on the idea that an intervention seeking to combat misinformation might
78 end up reinforcing 'in-group' thinking among those most conspiracy-minded or most politically polarised. For
79 these people, it is speculated that an intervention (e.g. labeling their favoured sources as false or
80 untrustworthy) could further entrench them in their distrust of legitimate public health messaging. This has
81 the potential to make the intervention not only less effective, but potentially negative in impact. This is
82 known as the 'backfire effect' and will be evaluated in the included studies. A concept arising from the policy
83 and psychology disciplines that could contribute to addressing potential backfire effects is framing. Framing
84 refers to the use of strategic messaging that is created with the intent of aligning with the extant worldview of
85 the target audience to make new ideas or information as congruent as possible. In practice, framing has been
86 found to improve fact-checking and accuracy nudge interventions [9].

87 There are studies testing interventions, and many reviews of the theory surrounding misinformation, but as
88 yet no reviews attempting to achieve a broader overview and evaluation of the various interventions which
89 emerged in the COVID-19 context. This project aims to contribute both toward the expansion and
90 application of realist review approaches, while simultaneously contributing toward better understandings of
91 interventions against COVID-19 misinformation. As COVID-19 continues to spread and the possibility of a
92 new pandemic lurks as an ever-present threat, developing the best understanding of interventions to
93 effectively combat COVID-19 misinformation will serve to help prepare policymakers and public health
94 apparatuses for the next pandemic.

95 **Research Question: Which interventions are most effective in combating spread of and belief in**
96 **COVID misinformation?**

97 **Sub-questions:**

98 **RQ1: Which types of interventions work best?**

99 **RQ2: Which groups of people do they work for?**

RQ3: Under which circumstances are the interventions most effective?

RQ4: What is the quality of studies testing interventions to combat spread of and belief in misinformation?

Results

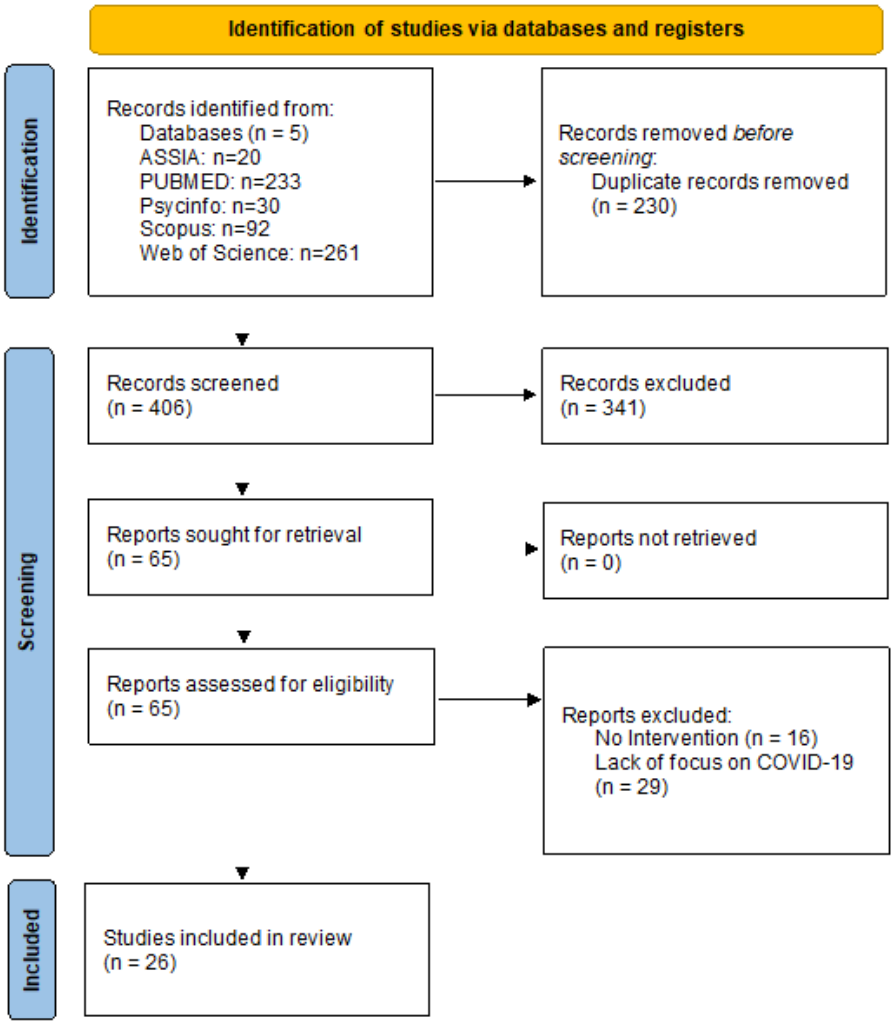
Study characteristics

26 papers were found that met inclusion criteria, including 6 out of reviewing the bibliographies of the 20 studies found through the search strategy described above. 636 were initially resulting from the searches, with 230 duplicate results removed, 341 deselected by title, and 45 deselected by full-text review, resulting in 26 eligible papers (Figure 1). Papers were published between 2020 and 2023, with a variety of national, regional, and international participant groups and study origination countries. The papers reviewed utilised participant groups coming mainly from the USA through private research participant companies like MTurk, Lucid, Prolific, Pollfish, and YouGov but also targeted audiences within the US like essential workers [10] and 'Latinx' communities [11]. Beyond the US, participant groups from Germany, the UK, Hong Kong, China, Canada, the Netherlands, Brazil, Kyrgyzstan, India, and internationally were included in the reviewed studies. These studies split into the intervention framework developed in the data extraction process as follows: 6 studies using Accuracy Nudges; 6 using education; 3 using Prebunking; 3 using Games; 3 using message framing; 3 using Community Engagement; and 2 using Debunking (Table 1). Full details of the studies can be seen in the appendices.

Figure 1. PRISMA flow diagram for systematic reviews

Table 1. Study characteristics table

PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only



Study	Intervention-type	Quality (/8)	Mechanism	Delivery Method	methodology	n=	Outcome
Amin et al. 2021	accuracy nudges	3	stimulating attention	Visual Selective Attention System tool	experiment	38	decision behaviour improved
Aslett et al. 2022	accuracy nudges	7	source credibility labels	embedded in feeds and search results	randomized field experiment	3337	no significant change
Dias et al. 2020	accuracy nudges	5	logo banner	trial, but presented in facebook format	survey experiment	6987	no, even potentially counterproductive
Gavin et al. 2022	accuracy nudges	4	accuracy reminder	online trial survey	replicating studies in other regions	2581	lowered willingness to share misinformation
Kreps et al. 2022	accuracy nudges	5	false tags	trial assignment modeled after Facebook	survey experiment	2000	little effect on veracity judgment or sharing
Pennycook et al. 2020	accuracy nudges	7	accuracy reminder	reminder at beginning of study	RCT	1709	accuracy nudges are simple and effective
DeGarmo et al. 2022	community engagement	6	community outreach	community health promoters	randomized control trial	1841	significant, medium size effect
Maragh-Bass et al. 2022	community engagement	3	digital storytelling	workshop developing them, then sharing	community workshops, storytelling	11	Suggests effectiveness for marginalised communities
Ugarte & Young 2023	community engagement	6	peer leaders	group chats within private facebook groups	two-arm, parallel-group, RCT	120	results suggest it lowers misinformation spread
Vijaykumar et al. 2021	debunking	5	corrective information	trial assignment	two mixed-design experiments	1454	enhanced trust and sharing of accurate information
Yousuf et al. 2021	debunking	5	debunking video	trial assignment to watch the video	randomized trial	980	significantly stronger rejection of misconceptions
Agley et al. 2021	education	7	infographic	viewing as part of the trial	two-arm, parallel-group, RCT	1017	small effect but highly scalable
Fung et al. 2022	education	4	educational phone call	telephone	multi-week educational intervention	25	significant educational improvements

Johnson et al. 2022	education	6	real social media	trial assignment to watch videos	RCT	842	significant success compared to a control
Van Stekelenburg et al. 2021	education	7	infographic	trial assignment	longitudinal survey	1202	did not significanly improve belief accuracy
Vandormael et al. 2021	education	7	educational video	social media distribution internationally	RCT	15163	effective at boosting preventative knowledge
Veletsianos et al. 2022	education	1	educational comic	trial assignment to read the comic	post-test only non-experimental design	295	Results indicate comic was effective and engaging
Basol et al. 2020	games	6	the game	trial	replication and extension experiment	196	significantly improves veracity judgment
Ma et al. 2023	games	4	the game	trial assignment	multi-study RCT	311	enhanced misinformation discrimination
Maertens et al. 2021	games	8	the game	trial assignment	longitudinal experiments	515	lasting increase in misinformation discernment
Bender et al. 2023	message framing	4	framing	physician presenting information via video	randmized 2x2 between-subject design	652	Small but significant impact
Freeman et al. 2021	message framing	6	framing	trial provision of written information	single-blind, parallel-group, RCT	15014	effective on the most vaccine-hesitant
Iles et al. 2022	message framing	8	framing	online trial assignment	randomized online experiment	1804	significant reduction in vaccine-hesitancy
Amazeen et al. 2022	prebunking	6	inoculation messages	self-administered online survey	inoculation messages	540	only among those with healthy attitudes
Jiang et al. 2022	prebunking	7	inoculation messages	trial assignment reading	3 phase between-subject experiment	123	generated superior resistance to misinformation
Piltch-Loeb et al. 2022	prebunking	5	inoculation messages	video	quasi-experimental, with control	1991	significant effects compared to control

All eligible studies underwent quality assessment using Kennedy et al.'s [12] risk of bias tool for assessing study rigor, results are shown in Appendices Table 1.1. Many studies lost several points due to lack of follow-up elements or not giving information on whether comparison groups were equivalent on demographics or baseline outcome measures. Iles et al. [13] and Maertens et al. [14] stand out as the only perfect scoring studies, with Veletsianos et al. [15] on the other side scoring only 1/8 as the lowest score of the assessed studies. When sorted into intervention-types, the average quality scores are relatively similar for each group, indicating a similar level of quality across the intervention-types.

Intervention characteristics

The studies in this review tested interventions with far greater heterogeneity than the dominant interventions proposed before the COVID-19 pandemic (accuracy nudges and fact-checking). As can be seen above in the study characteristics table, the studies were iteratively sorted into intervention-types as laid out in the methodology section. These intervention-types included: accuracy nudges, community engagement, debunking, prebunking, education, games, and message framing. This section will briefly introduce these intervention types and their defining characteristics.

Accuracy nudges in the reviewed studies consisted of mechanisms including: stimulating attention [16], source credibility labels [17], logo banners to help identify trustworthiness of sources [18], accuracy reminders [4], and tags that mark information as false [19]. These various intervention mechanisms fit under accuracy nudges due to their common characteristics as simple, fast, attention-grabbing labels or reminders that 'nudge' the participant to consider information veracity and bring that consideration into the forefront of their minds immediately before reading the information.

Community engagement is difficult to characterise by intervention mechanism because the defining aspect of community engagement occurs *before* intervention mechanism is determined in the research design. Instead of pre-determining intervention mechanisms and delivery methods, community engagement involves co-creation of the intervention alongside and in collaboration with the targeted community, to be bespoke to the unique context and circumstances of the community [10, 11, 20].

Debunking refers broadly to reactive interventions (e.g. fact-checking) that seek to 'debunk' existing misinformation and help people exposed to it rethink their belief and formulate new understandings of the relevant information [21, 22]. In contrast, 'prebunking' interventions seek to build resilience to misinformation in people preemptively before exposure has occurred, and potentially even before the piece of misinformation has been created/spread [23]. This typically takes the form of inoculation messages administered to participants before exposure to potential misinformation. In this way they are similar to accuracy nudges – the key difference being that prebunking is more extensive than accuracy nudges. The inoculation messages are more significant, take longer to process, and are intended to take the full attention of the participant for the duration of the message, whereas accuracy nudges are fast and often involve the periphery of a participant's attention. In the reviewed studies characterised as prebunking, all three involve inoculation messages as their intervention mechanism [24, 25, 26].

Education is the most heterogeneous of the intervention-types and can be difficult to categorise as educating the participant is essential to all interventions working to address misinformation. In the reviewed studies, this intervention-type involved mechanisms such as: videos [27], comics [15], infographics [28, 29], and a multimodal intervention using authentic social media messaging [30]. The defining characteristic of the reviewed studies in this intervention-type is the primacy and exclusivity of education as the goal of the intervention. For instance, in Vandormael et al. [27], an educational video was released and distributed internationally with the goal of maximising viewership, but with no additional features of the intervention beyond watching the video.

Game intervention-types are characterised by the inclusion of a computer game for participants to play as the primary intervention-mechanism. This can be seen in all three of the included studies under this categorisation. These games inform players (participants) on the tactics and manipulation used to create and spread misinformation, with the goal of creating an inoculation effect and helping bolster veracity-judgment in participants. For example, Bad News is the name of the game used in Maertens et al. [14], a popular game used in many studies outside the purview of this review as well. In this game, players take on the role of an antagonist, creating misinformation and working to spread it through social media and the internet.

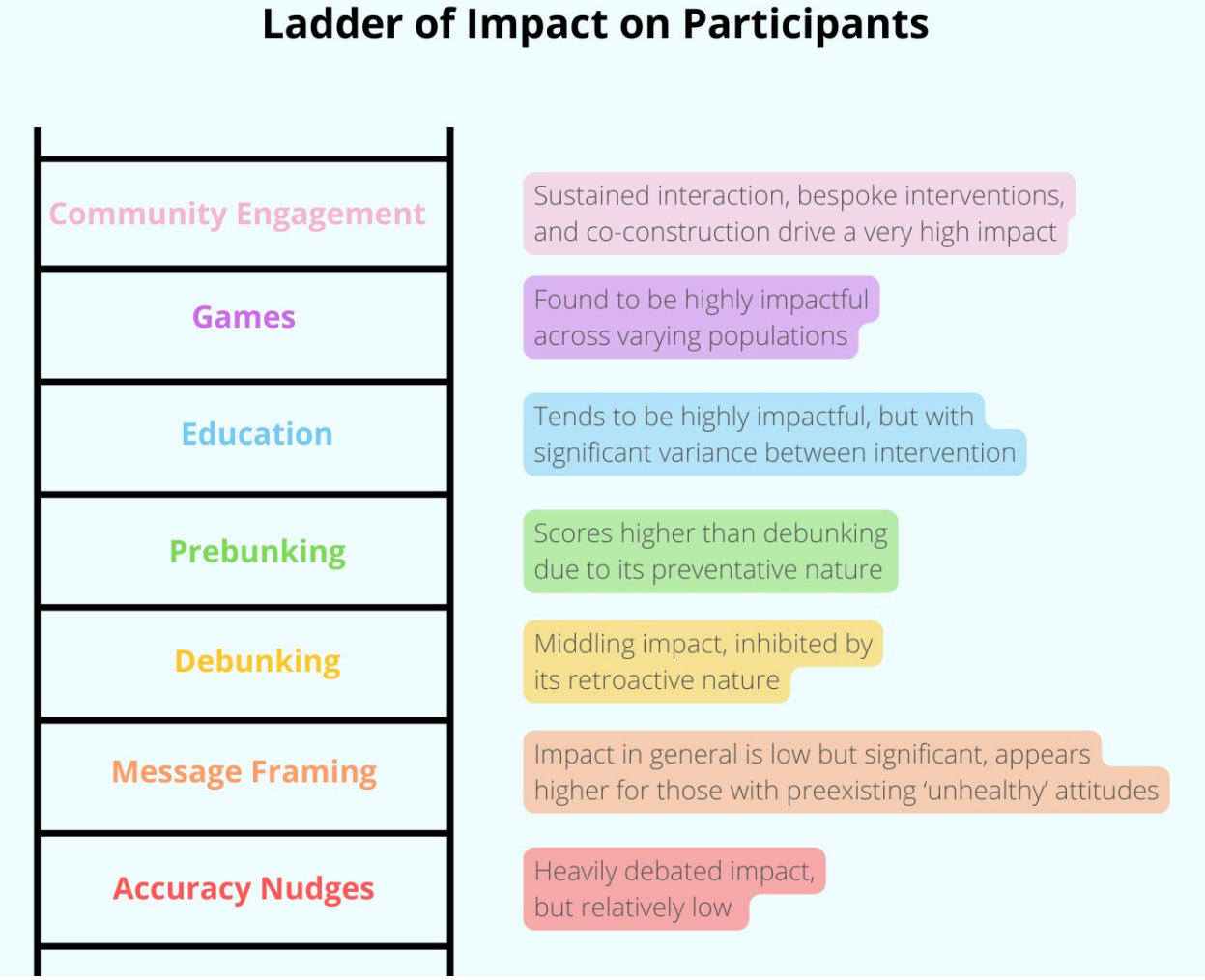
Message framing as an intervention-type is characterised by the use of psychological framing in the development of the language used in the intervention. Whether presenting information via video or written information, what distinguishes these studies as message framing is their strategic use of language to attempt to make their information transfer to participants as congruent with their extant worldview as possible. This then helps participants internalise that information effectively and can address intervention design concerns around potential backfire effects.

Intervention effectiveness

The two variables most central to answering which interventions work 'best' appear to be scalability and impact. If impact is too low, the intervention might not actually engender sufficient behavioural change in participants to combat the misinformation. Similarly, if an intervention cannot be upscaled, it has no capacity to address COVID-19 misinformation at a systemic level. The 'ladder' visuals below represent the intervention-types relative to one another across these two variables (Figures 2 & 3). Relative impact is determined by the measured impact on participants in each study. These measures are not consistent, yet with the authors' interpretations, comparisons are possible. These relative measurements focus on the impact per participant, with no regard to number of participants or scalability. Inversely, the scalability ladder visual focuses on scalability with no regard to impact per participant. These visuals are meant to simplify and ease understanding of the results, and are purely relative.

Figure 2: Ladder of Impact on Participants

Figure 3: Ladder of Scalability



On the bottom rung of impact per participant is accuracy nudges, whose impact is heavily debated. Some authors in this review such as Pennycook et al. [4], champion this intervention type and claim significant impact in their results. Gavin et al. [8], who replicated Pennycook et al. [4], found mixed results that stood at odds with the original study. Amin et al. [16] found impact on decision behaviour and tendency to share misinformation with their study, but the rest of the studies in this intervention group found either minimal impact [19], only impact on certain groups [17], or no impact [18] who even noted potential counterproductivity.

Framing of public health messages is next along the impact ladder. Studies testing interventions using different framings of public health messaging found significant impact [13, 31, 32], although not as high as

other intervention-types included in this review. This impact was largely reserved for those heaviest consumers of misinformation and those most vaccine-hesitant [31].

Debunking by its nature must occur retroactively, which limits impact as the initial exposure must be overcome. In this way, debunking has two independent goals: to both disprove internalised misinformation and convince the participant of the veracity of legitimate information. This is a barrier to impact, which is noted by both Vijaykumar et al. [21] and Yousuf et al. [22]. Vijaykumar et al. [21] found no impact on perception or willingness to share misinformation yet found enhanced credibility and readiness to share accurate information because of their intervention. However, Yousuf et al. [22] found that exposure to their intervention did result in enhanced trust in government and significantly stronger rejection of vaccination misconceptions.

Prebunking, as the preventative version of debunking, scores better on impact. Prevention is found to be more powerful in a variety of aspects than reactive debunking. All three included studies [24, 25, 26] found significant impact among participants, although in the case of Amazeen et al. [24] this significance was limited to those with preexisting 'healthy' attitudes. Impact on participants was found to include generating resilience against misinformation, less willingness to share misinformation, and greater willingness to receive a vaccine.

Education is the most varied type of intervention with a range of impact between the individual tested interventions (the relative score here is an aggregate). At best, educational interventions have the potential to be a form of systematic prebunking with great effectiveness. In the reviewed studies, they were found to improve knowledge and increase resilience to misinformation at significant levels, particularly among populations with low preexisting knowledge levels [27, 30]. However, Van Stekelenburg et al. [29] found no significant impact, highlighting the variability of this intervention-type.

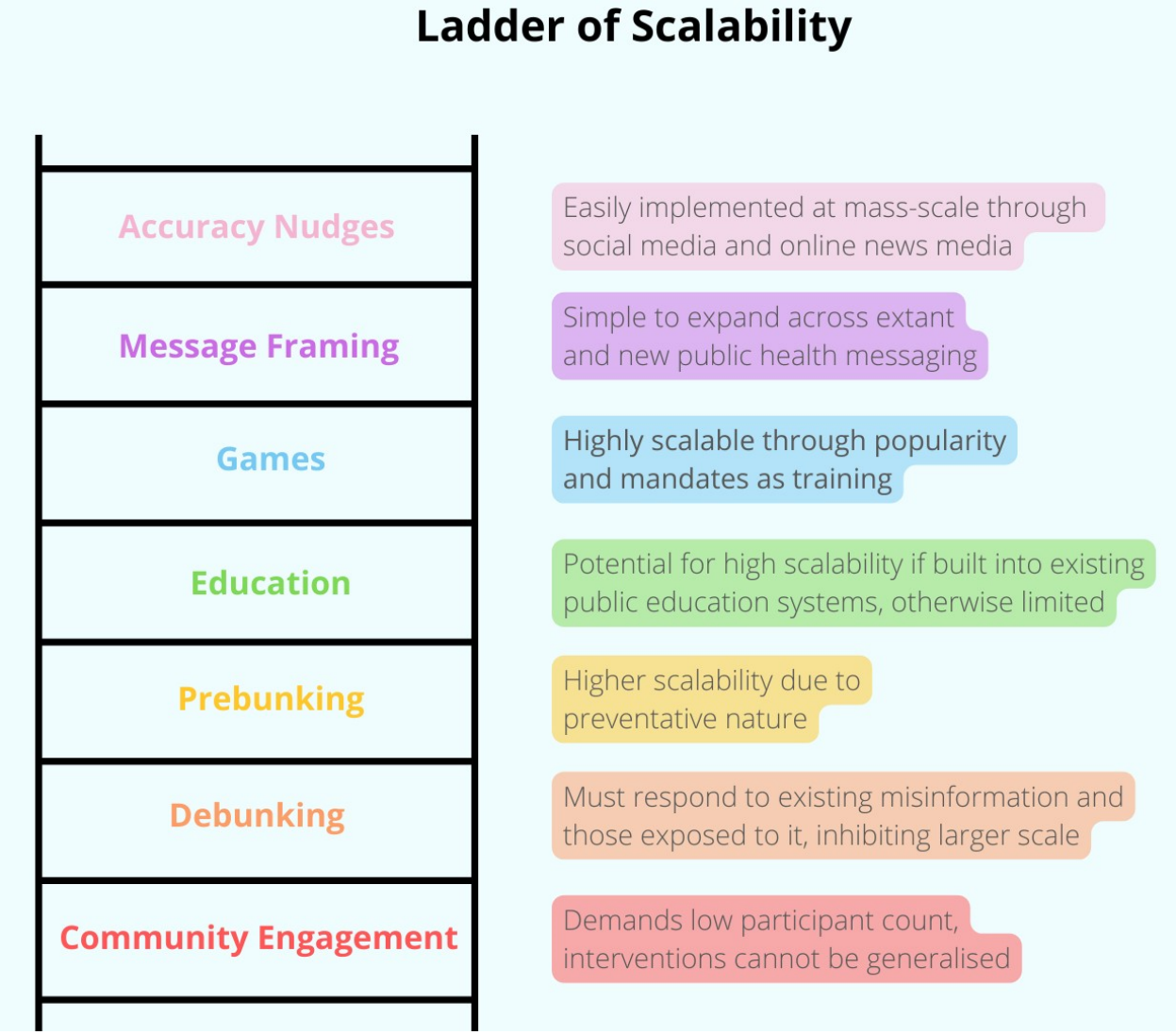
Games were consistently found to be highly impactful across the various populations who played them, with high levels of durability and longevity compared to other intervention-types reviewed and significant impact levels for all kinds of preexisting attitudes towards vaccination and COVID-19 [14, 23, 33]. Every reviewed

227 study found significant impact, which corresponds with a high relative impact score, although still below the
228 bespoke and prolonged interventions within community engagement.

229 Community engagement is the single most impactful intervention-type, with sustained interaction and
230 bespoke interventions to specifically targeted communities who then themselves are brought into the
231 intervention process and invited to participate, make their voices heard, and have their concerns addressed in
232 a bespoke, personal, and trusted manner. All reviewed studies found significant and extensive impact among
233 their participants.

234

235



Community engagement is essentially impossible to scale upwards. It inherently requires small numbers of participants and high levels of resource and time investment by those implementing the intervention. The interventions themselves are then not even intended to be generalisable, but rather bespoke to and befitting the contextual needs of the community involved. Community engagement can only effectively be done at a small scale over long periods of time involving the building of trust with community, the proactive engagement and co-creation of interventions and implementation strategies with the community itself, and the implementation strategies themselves can take years to accomplish [10, 20].

Debunking scores quite low in scalability. Debunking at a large scale is extremely difficult as it is inherently based on preexisting misinformation and cannot effectively prevent additional misinformation. Further, it must always attempt to reach those specific populations initially exposed to the targeted misinformation to be 'debunked', which is difficult and resource-intensive.

Prebunking does not need to attempt to find and target those who already saw some misinformation as the intervention occurs before misinformation is seen. For this reason, prebunking is easier to scale upwards than debunking and is relatively lower in resource-cost. Implementation of prebunking involves the development of 'inoculation messages' [24, 25] as written messages or video content intended to raise resiliency of participants against misinformation.

Without being built into the public education system, educational interventions may struggle to scale upwards, relying on peer educational champions [34] or social media 'virality' to spread [27]. Adjusting anything within the public education system is highly resource-intensive, even though those changes are then highly impactful and wide-reaching. However, when performed in smaller scale as in the included interventions, educational interventions can be substantially reduced in resource intensity [27].

Although not as easily scalable as message framing or accuracy nudges, games are nonetheless relatively highly scalable when compared to the other intervention-types in this review. As the games are already developed, introducing them to new populations is then relatively simple, resource-inexpensive, and quick.

Message framing has high scalability with the simple addition of language strategising and purposeful narrative framings applied to extant and new public health messaging. Message framing is only slightly more resource intensive than accuracy nudges in that it must be bespoke to particular narratives, communities, and groups. However, in each bespoke circumstance, still the resource intensity would be low.

Accuracy nudges is undeniably the highest scalability intervention-type. The core reason why accuracy nudges are so scalable is the extremely low resource intensity needed to implement them. It requires the insertion of nudges in social media feeds and news articles. This would be easy and inexpensive for social media

corporations and newspapers to implement, even when scaled into the extreme levels of interaction and users involved in contemporary social media.

Durability of effect

In the studies that did test for longevity/durability of impact in their tested intervention, consistent low levels were found, with findings indicating high reliance on intervention repetition and regular testing of misinformation resilience over a sustained (and potentially indefinite) period to reach functional durability of effect. The study that looked most closely at this was Maertens et al. [14] which performed one of the only longitudinal studies included in this review explicitly investigating longevity of impact using the 'Bad News' game as its chosen intervention. They found that their intervention resulted in a significant increase in ability to discern misinformation with lasting effects if regular misinformation resilience testing occurred over time. Without regular testing they found significant decay over a 2 month period ending in a loss of inoculation effect [14].

Special groups and circumstances

There appears to be a significant distinction in how these interventions work between those with preexisting 'healthy' understandings of public health information and those who are the heaviest consumers of misinformation. This was noted in several studies ([17, 18, 19, 23, 24, 28, 32], and in ways that do not initially appear congruent with one another. It is clear this subgroup of heaviest misinformation consumers is impacted differently by many of the interventions included in this review, but that change in impact is not a consistent factor - instead it is an ephemeral variable, difficult to spot and even harder to plan for in study design.

The table below lays out the contexts in which each relevant intervention type was found to be most effective in the groups tested, alongside the groups included within the included studies, relevant findings from the authors regarding context and their intervention, and an overall level of generalisability (Table 2).

Table 2. Intervention context and generalisability

Intervention	Contexts for use	Pilot groups within the studies	Findings on context	Generalisability
Accuracy Nudges	Internet, social media platforms	1. USA [YouGov] proportional group online (Aslett et al. 2022) 2. Students in 'individual chats' on Android devices (Amin et al. 2021) 3. USA [MTurk] group online (Dias et al. 2020) 4. Kyrgyzstan, India, and USA non-probability samples (Gavin et al. 2022) 5. USA [Lucid] proportional group online (Kreps et al. 2022) 6. USA [MTurk] group online (Pennycook et al. 2020)	1. Intervention only works on heaviest misinformation consumers 3. Potential for 'Backfire' effect on those most misinformed 5. Intervention effectiveness changed with location 6. No evidence of 'Backfire' effect	Very high
Prebunking	Universally	1. USA [YouGov] proportional group online (Amazeen et al. 2022) 2. Hong Kong undergraduate students (Jiang et al. 2022) 3. USA [Pollfish] proportional group online (Piltch-Loeb et al. 2022)	1. Intervention only works on those with preexisting healthy attitudes	High
Games	Youth, digitally literate people, employment mandates	1. USA [Prolific] proportional group online (Basol et al. 2020) 2. China [WeChat] group online (Ma et al. 2023) 3. USA [Prolific] proportional group online (Maertens et al. 2021)	1. Intervention works across the political spectrum 2. Intervention works well for general public	High
Debunking	Reactively to widely believed	1. UK and Brazil Whatsapp users (Vijaykumar et al. 2021)	1. Most effective on older people	Medium

	misinformation, older people	2. Netherlands elderly (Yousuf et al. 2021)		
Education	Difficult to reach communities, communities distrustful of government where peers and individual study might be most effective	1. USA [Prolific] proportional group online (Agley et al. 2021) 2. USA [MTurk] group online (Johnson et al. 2022) 3. Older adults in Hong Kong (Fung et al. 2022) 4. USA [Prolific] proportional group online (Van Stekelenburg et al. 2021) 5. Canada/USA [Prolific] women online (Veletsianos et al. 2022) 6. International social media users (Vandormael et al. 2021)	2. Intervention worked best on older and less vaccine-hesitant people 4. Intervention caused 'Backfire' effects in conservative Republicans 6. Most effective for low baseline knowledge levels	Medium
Message framing	distrustful communities, those 'bought-in' to conspiracy already, when dealing with political polarisation	1. Germany non-probability sample (Bender et al. 2023) 2. US [MTurk] group online (Iles et al. 2022) 3. UK [Lucid] proportional group online (Freeman et al. 2021)	1. Extant framing best for those anxious about vaccination. Intervention framing best for those strongly anti-vaccine 2. Emphasising personal benefit more effective on those most vaccine hesitant. 3. Emphasising collective benefit creates 'Backfire' effects	Low
Community Engagement	deprived communities, vulnerable communities, outliers	1. American 'Latinx' communities (DeGarmo et al. 2022)	1. Effective for mitigating health disparities	Very Low

		2. USA young black adults (Maragh-Bass et al. 2022) 3. USA 'essential workers' (Ugarte et al. 2023)		

Eight reviewed studies found insignificant trends in intervention impact between baseline participants and special groups, with several more looking for such trends and finding none. This indicates the specificity of these intervention-types, and that although context and social group could be determinants of intervention effectiveness, such effects are likely to be small. For example, Bender et al. [32] noted that their intervention framing worked best on those already strongly anti-vaccine. Conversely, Johnson et al. [30] found their intervention worked best on those with less vaccine hesitancy, and that those with higher social political conservatism performed worse on knowledge scores. The insignificant trends found in these studies were typically tied to either age, ethnic group, or political ideology as core identities tied to perceptions and experiences of COVID-19 and the public health responses thereto. Political (rightwing/conservative) ideology was noted in many studies as a subgroup of particular importance and was found to coincide with less accurate pre-intervention beliefs [23, 28].

Accuracy nudges were tested with participants from the USA, Kyrgyzstan, and India, with findings that suggest that their impact is difficult to predict and changes depending on the context [8]. Dias et al. [18] noted the potential for a 'backfire' effect among those people most bought-in to misinformation, whereas Kreps et al. [19] found no evidence of this effect. Aslett et al. [17] find that their intervention only worked on those who consume the highest levels of misinformation in their participant group and had minimal effect on anyone else. This conflicts with concerns about backfire effects.

Prebunking was tested with participants from online recruiters in the USA and Hong Kong undergraduates. Amazeen et al. [24] found that the intervention only worked on those with preexisting 'healthy' attitudes, meaning those whose beliefs already coincided most closely with legitimate public health messaging. Because

this intervention-type is intended to inoculate the 'average' person against misinformation, it only working on those with preexisting 'healthy' attitudes does not reduce the usefulness of prebunking.

Games were tested in the USA and China with proportionally representative online groups. Basol et al. [23] as well as Ma et al. [33] respectively found that the interventions worked across both the political spectrum and the public in general. This indicates high generalisability, particularly with the proportionally representative and relatively large participant cohorts in these studies. However, by the nature of a digital intervention type like games, older people and those with low levels of digital literacy (who are among those most desirable to target for the intervention) may have less desire or ability to play the game.

Debunking was tested in the UK and Brazil among Whatsapp users, and in the Netherlands among the elderly. Interestingly, Vijaykumar et al. [21] found that their intervention was most effective on older people. This indicates that this type of intervention might be most useful among elderly populations and communities. Vijaykumar et al. [21] and Yousuf et al. [22] speculate that perhaps older people have higher baseline trust in governmental messaging and are therefore more open to changing their internalised beliefs based on new information from legitimate sources. By the nature of debunking, it can only be applied reactively to widely believed misinformation, which significantly limits its generalisability.

Education was tested in the USA, Canada, Hong Kong, and internationally through social media sharing. Johnson et al. [30] found their intervention worked best on elderly people and those with less hesitancy around COVID-19 vaccination. Similarly, Veletsianos et al. [15] found that their intervention caused a noteworthy 'backfire' effect among conservative US republicans (as the most vaccine-hesitant and 'bought-in' to misinformation already). Vandormael et al. [27] suggested educational interventions might be most effective among populations with a low baseline knowledge level, as their own participant group has relatively high levels of baseline knowledge (although nonetheless the intervention successfully boosted knowledge of COVID-19 prevention). When taken together, these findings indicate that the groups most ideal for this type of intervention are communities with low baseline knowledge of public health information or communities distrustful of government where peer and individual study might be able to penetrate that distrust.

Message framing was tested in Germany [32], the US [13], and the UK [31] all through online interventions testing framed messaging against traditional extant public health informative messaging. Bender et al. [32] found that extant framing (which typically focuses on collective benefits and informing about vaccination side-effects) worked best for those anxious about vaccination, whereas the intervention framing worked best for those strongly anti-vaccine. Similarly, Freeman et al. [31] found that emphasising personal benefit (the intervention framing) was more effective on the most vaccine-hesitant, whereas emphasising collective benefit (the control/extant framing) was far less effective and even resulted in 'backfire' effects. Together these findings make a strong case for message framing interventions to effectively target those communities most distrustful of government messaging, those most 'bought in' to conspiracy and misinformation already, and the most politically radicalised.

Community engagement was tested in the US among 'Latinx' communities [11], young Black adults [20], and 'essential workers' [10]. By its nature, community engagement is very low generalisability as it is more contextually specific, resource intensive, and time-consuming than any other intervention type. Degarmo et al. [11] found their intervention was successful at mitigating health disparities in the communities they engaged. This suggests community engagement would be most effectively utilised in deprived communities, vulnerable communities, and those areas most difficult to reach for any reason.

Discussion

The research questions in this study do not have explicit ranked answers, as impact and scalability differ widely across the interventions included in this review. There are tradeoffs in play, between impact and scalability as well as between generalisability and targeted intervention against subgroups of particular importance. Therefore, the key finding from this review is the insufficiency of any one intervention to address the widely varying needs of the many contexts and groups in which misinformation can spread. There is a need for the development of comprehensive packages (each containing multiple interventions) as the core policy recommendation. These packages can pull from the different strengths of each intervention type reviewed to best fit the needs of the relevant communities and contexts within which these packages will be

developed. When such a package of multiple interventions is impossible, game-type interventions appear to be an outlier in terms of being highly scalable, impactful, low resource-intensity, and highly generalisable relative to the other intervention-types reviewed.

Politics and partisan bias

Both the theoretical and intervention literatures around COVID-19 misinformation hint at its politically polarising elements yet fail to address this influence head on. Dispersed throughout the findings and discussions of the included studies are the political elements of COVID-19 misinformation. It is consistently found that political conservatives, particularly in the US, are uniquely vulnerable and bought-in to misinformation and conspiracism [7, 35]. This group was found to have its own unique interactions with many of the tested interventions in this review. When this happened, the authors mention this difference and give some speculation as to why that might be the case, but do not investigate this finding further, or seek to use explanations in the wider literature to support their findings (see [29] for the most comprehensive discussion of this issue in the eligible studies). Additionally, there has been very little work to explicitly begin from this starting point and deep dive into why this might be the case and how interventions might most effectively impact this group. This presents a significant detriment to reaching the stated goal of these interventions - effectively combatting COVID-19 misinformation.

Pennycook et al. [4] is the most influential study included in this review in terms of citation count, references throughout the reviewed studies, and the extent to which their study has been replicated and critiqued within both the studies under review and the wider literature. Within that study they champion the theory that the systemic sharing behaviour of COVID-19 misinformation in our society is "because [people] simply fail to think sufficiently about whether or not the content is accurate when deciding what to share" [4, p. 770]. Pennycook et al. claim that their findings and this theory indicate that accuracy nudges are not only simple and effective, but the only intervention needed against COVID-19 misinformation. In doing so, they negate the claims of many of the other included studies in this review. This has brought significant criticism against this core idea of what is causing vulnerability to COVID-19 misinformation. If the only issue is a lack of

thinking, then accuracy nudges are the obvious intervention. Yet although the findings of Pennycook et al. [4] do suggest the effectiveness of accuracy nudges and the need for interventions that make people think more about their sharing decisions, this 'theory' they promote is insufficiently supported when applied to negating the findings of other studies. Their findings suggest the effectiveness of accuracy nudges, but not the ineffectiveness of other interventions. The alternative proposed answer to what is causing vulnerability to misinformation is partisan bias. This explanation posits that it is not failing to think sufficiently or lower cognitive ability that leads to vulnerability to misinformation, but rather the inherent bias that arises from adherence to political ideology in the context of intense political division and polarisation as is affecting the contemporary United States very deeply but also affects many countries today [36]. This debate on partisan bias vs insufficient thinking punctuates the literature on misinformation, including many of the studies included in this review.

Limitations

A primary limitation in this review comes from the heterogeneity of the studies and interventions disallowing meta-analysis and other forms of traditional systematic review analysis that rely on similar outcome measures and methodologies within the eligible studies. This limitation is accentuated by the potential for interpretation bias. The interpretation of the data herein is biased by the perspective and worldview of the authors. Additionally, there is limited consistency between realist reviews and limited standards and assessments available to apply to this review. This does not necessarily limit the rigor of the review but makes analysing that rigor and validity more difficult. The development of more and consistent direction and assessments for realist reviews would address this limitation currently present within the method. Lastly, the limited engagement in the intervention literature with theory limits the extent to which theoretical insights can be drawn from this study.

Future Research Directions

Although a variety of interventions tested in the studies herein found success in the short term, in the long-term it is impossible to avoid the urgent need for mass-scale education on digital literacy if the goal is to make

a population as resilient as possible against misinformation. Future research in this direction is pivotal, with experiment-groups in classrooms a clear next step. Additionally, future research on how to address the political difficulties in implementing such a wide-scale intervention is required.

Out of all intervention-types reviewed, games appear to create the highest impact while still being highly scalable and resource-inexpensive, with the potential for longevity in the right conditions [14]. Relative to the other intervention-types, games scores maximally in terms of impact on participants, while still being relatively high on scalability. Future research in this direction is needed to refine and test these results. Longitudinal testing is an obvious follow-up to gain insight into durability of inoculation effect.

Additional areas for future research include: 1) theoretical research into how to build a resilient population and how to address vulnerability to misinformation systemically versus individually; 2) the role of politics and partisan bias in the functioning of these interventions; 3) where misinformation comes from and who gains from it; 4) the role of political polarisation and radicalisation in vulnerability to and the spread of misinformation.

Conclusions

This review included 26 studies of interventions combatting COVID-19 misinformation. The interventions reviewed varied widely in terms of scalability, resource intensity, impact on participants, the contexts within which each best works, the people onto whom the interventions will have greatest effect, and research quality. The tests performed in the included studies hold rich contributions toward better understanding how misinformation functions, how veracity judgement occurs in individuals and communities, and which interventions work best in which contexts and for whom. COVID-19 showed precisely how harmful and deadly misinformation can be, and what a public health threat it can represent. In this fight against systemic misinformation in our society, a final takeaway from this review is the need for acknowledgement of misinformation as a societal and systemic issue that requires significant investment and time to resolve, if resolution is possible.

Materials and methods

This review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses [37] checklist, available in the appendices. The review followed a pre-registered protocol submitted before the study began with PROSPERO, registration number CRD42023440580, record title: "Realist review: assessing intervention effectiveness in combating COVID misinformation", available at the PROSPERO website. Amendments to the information provided in the protocol were centred on the elimination of an initially-planned research question on the intersection of theory and intervention literatures within the reviewed studies. This research question was removed after data extraction and analyses revealed a dearth of theoretical investigation in the reviewed studies. Instead this lack of theoretical involvement in the reviewed studies is noted in the discussion.

Search strategy

This review included a systematic search of Web of Science, Scopus, ASSIA, Psycinfo, and Pubmed to identify English language articles written between January 1, 2020 and June 22, 2023 performed following a pre-registered protocol conforming to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement [37]. No secondary searches were performed. Search strategy followed the protocol using pre-determined search terms, with results imported into Excel sheets for ease of deselection. Duplicates were removed and then an initial title-based screening was performed. Screening then followed based on abstract and then full-text review. Additional searching among the references of the included studies followed. Duplicate screening was performed by a team member (K.G.) on ~15% of studies through all screening stages, with any disagreement resolved via discussion. Inter-rater agreement was found to be very high (~92%).

The full search-string chosen for this review, which was only applied to Titles and Abstracts, is as follows: (conspirac* OR anti-vax* OR anti-vaccine OR 'anti vaccine' OR misinform* OR fake OR fals*) AND

(messag* OR rumor* OR argu* OR rhetoric OR spread*) AND (COVID OR COVID-19 OR coronavirus OR 'corona virus' OR pandemic*) AND interven*

Eligibility

Trials or experimental studies were eligible if they were focused on reducing the spread of and vulnerability to COVID-19 misinformation in their participants, and tested an intervention meant to combat COVID-19 misinformation. Studies were required to be in the English language and been published between 2020 and 2023 as searching before the COVID-19 pandemic began was unnecessary.

Quality assessment

The methodological quality of each study chosen for inclusion was assessed via Kennedy et al.'s [12] risk of bias tool for assessing study rigor. It includes eight items for appraisal: (1) cohort, (2) control or comparison group, (3) pre-post intervention data, (4) random assignment of participants to the intervention, (5) random selection of participants for assessment, (6) follow-up rate of 80% or more, (7) comparison groups equivalent on sociodemographics, and (8) comparison groups equivalent at baseline on outcome measures. This assessment tool was used for its flexibility regarding type of methods and interventions in the studies being assessed. Although this analysis was performed, no studies were excluded due to quality as realist reviews explicitly disagree with exclusion from quality concerns as explained below.

Data extraction and analyses

The following information from included studies was extracted into a table to highlight study characteristics as can be seen in the next section: Study, intervention-type, 'working ingredient', 'delivery method', country of origin, methodology, number of participants, and whether the intervention was found to be successful. Additionally, a variety of other information was extracted to inform the other tables and charts found in the results section. All text from the eligible studies was imported into NVivo Pro 14 and the methods, results, and discussion sections underwent qualitative coding. Coding was done iteratively to categorise the findings. This iterative process evolved into a developed framework as the coding took place. For instance, if one

intervention was identified from an article during coding, the coder attempted to assign it to a category within the emerging intervention framework. New subcategories were created if the current categories were insufficient, until all interventions were categorised. As coding progressed, the intervention framework came to be populated through the included studies. The heterogeneity of the included studies and their respective measures disallowed quantitative meta-analysis.

Regarding effectiveness, impact per participant and scalability were the primary variables analysed. Impact per participant refers to the level of individual behavioural change experienced by the participants of each intervention reviewed, as all were centred on individual behaviour. Scalability is a more complex variable consisting of several combined factors including generalisability (how effectively can results be replicated in other contexts and with other groups), resource-intensity (how expensive is the intervention in terms of time, money, and overall resource expenditure), and capacity for upscaling (how many people it could reach). With impact and scalability thus defined, effectiveness can be then analysed by how many people could be impacted and to what extent per person. A sub-analysis of context was undertaken by comparing context by intervention type, and laying out which participant groups were targeted by the interventions. Additional analysis was performed to investigate context beyond the community of participants within the intervention.

It is important to define 'circumstances' as used in RQ3. Here, circumstances refers to the context within which an intervention is taking place - such as geographic location, identities and wealth of the targeted community, and structural and institutional factors within which the community and intervention will take place. Additionally, circumstances refers to the experience, resources, and capacity of the research team or implementing body performing the intervention.

Acknowledgments

We want to thank Katie Goddard from the Primary Care and Public Health department of the Brighton and Sussex Medical School for her help in duplicating and affirming the deselection and qualitative coding in this study.

References

1. Chan, M. 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
2. Cook, J., Ecker, U., & Lewandowsky, S. (2015). Misinformation and How to Correct It. In R. A. Scott & S. M. Kosslyn (Eds.), Emerging Trends in the Social and Behavioral Sciences (1st ed., pp. 1–17). Wiley. https://doi.org/10.1002/9781118900772.etrds0222
3. Cohen, D. (2023, January 19). Twitter Extends Community Notes to Quote Tweets. https://www.adweek.com/media/twitter-extends-community-notes-to-quote-tweets/
4. 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
5. Rathje, S., Roozenbeek, J., Traberg, C. S., Van Bavel, J. J., & van der Linden, S. (2022). Letter to the editors of Psychological Science: Meta-analysis reveals that accuracy nudges have little to no effect for US conservatives: regarding Pennycook et al.(2020). https://psyarxiv.com/945na/download?format=pdf
6. Lees, J., McCarter, A., & Sarno, D. M. (2022). Twitter's disputed tags may be ineffective at reducing belief in fake news and only reduce intentions to share fake news among Democrats and Independents. Journal of Online Trust and Safety, 1(3). https://www.tsjournal.org/index.php/jots/article/view/39

7. Gawronski, B., Ng, N. L., & Luke, D. M. (2023). Truth sensitivity and partisan bias in responses to misinformation. Journal of Experimental Psychology: General. https://psycnet.apa.org/record/2023-58364-001
8. Gavin, L., McChesney, J., Tong, A., Sherlock, J., Foster, L., & Tomsa, S. (2022). Fighting the Spread of COVID-19 Misinformation in Kyrgyzstan, India, and the United States: How Replicable Are Accuracy Nudge Interventions? https://assets.pubpub.org/d6zy4hsf/21663782255673.pdf
9. Featherstone, J. D., & Zhang, J. (2020). Feeling angry: The effects of vaccine misinformation and refutational messages on negative emotions and vaccination attitude. Journal of Health Communication, 25(9), 692–702. https://doi.org/10.1080/10810730.2020.1838671
10. Ugarte, D. A., & Young, S. (2023). Effects of an Online Community Peer-support Intervention on COVID-19 Vaccine Misinformation Among Essential Workers: Mixed-methods Analysis. Western Journal of Emergency Medicine, 24(2), 264.
11. DeGarmo, D. S., De Anda, S., Cioffi, C. C., Tavalire, H. F., Searcy, J. A., Budd, E. L., McWhirter, E. H., Mauricio, A. M., Halvorson, S., & Beck, E. A. (2022). Effectiveness of a COVID-19 testing outreach intervention for Latinx communities: A cluster randomized trial. JAMA Network Open, 5(6), e2216796–e2216796.
12. Kennedy, C. E., Fonner, V. A., Armstrong, K. A., Denison, J. A., Yeh, P. T., O'Reilly, K. R., & Sweat, M. D. (2019). The Evidence Project risk of bias tool: Assessing study rigor for both randomized and non-randomized intervention studies. Systematic Reviews, 8(1), 3. https://doi.org/10.1186/s13643-018-0925-0
13. Iles, I. A., Gaysynsky, A., & Sylvia Chou, W.-Y. (2022). Effects of Narrative Messages on Key COVID-19 Protective Responses: Findings From a Randomized Online Experiment. American Journal of Health Promotion, 36(6), 934–947. https://doi.org/10.1177/08901171221075612
14. 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.

15. Veletsianos, G., Houlden, S., Hodson, J., Thompson, C. P., & Reid, D. (2022). An Evaluation of a Microlearning Intervention to Limit COVID-19 Online Misinformation. Journal of Formative Design in Learning, 6(1), 13–24. https://doi.org/10.1007/s41686-022-00067-z

16. Amin, Z., Ali, N. M., & Smeaton, A. F. (2021). Visual Selective Attention System to Intervene User Attention in Sharing COVID-19 Misinformation (arXiv:2110.13489). arXiv. http://arxiv.org/abs/2110.13489

17. Aslett, K., Guess, A. M., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. Science Advances, 8(18), eabl3844. https://doi.org/10.1126/sciadv.abl3844

18. Dias, N., Pennycook, G., & Rand, D. G. (2020). Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. https://dspace.mit.edu/bitstream/handle/1721.1/144236/V2_researcharticle_publishers_jan29.pdf?sequence=2&isAllowed=y

19. Kreps, S. E., & Kriner, D. L. (2022). The COVID-19 infodemic and the efficacy of interventions intended to reduce misinformation. Public Opinion Quarterly, 86(1), 162–175.

20. Maragh-Bass, A., Comello, M. L., Tolley, E. E., Stevens Jr, D., Wilson, J., Toval, C., Budhwani, H., & Hightow-Weidman, L. (2022). Digital storytelling methods to empower young Black adults in COVID-19 vaccination decision-making: Feasibility study and demonstration. JMIR Formative Research, 6(9), e38070.

21. Vijaykumar, S., Jin, Y., Rogerson, D., Lu, X., Sharma, S., Maughan, A., Fadel, B., de Oliveira Costa, M. S., Pagliari, C., & Morris, D. (2021). How shades of truth and age affect responses to COVID-19 (Mis) information: Randomized survey experiment among WhatsApp users in UK and Brazil. Humanities and Social Sciences Communications, 8(1). https://www.nature.com/articles/s41599-021-00752-7

22. Yousuf, H., van der Linden, S., Bredius, L., van Essen, G. T., Sweep, G., Preminger, Z., van Gorp, E., Scherder, E., Narula, J., & Hofstra, L. (2021). A media intervention applying debunking versus

non-debunking content to combat vaccine misinformation in elderly in the Netherlands: A digital randomised trial. EClinicalMedicine, 35. https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00161-9/fulltext

23. 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). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6952868/
24. Amazeen, M. A., Krishna, A., & Eschmann, R. (2022). Cutting the Bunk: Comparing the Solo and Aggregate Effects of Prebunking and Debunking Covid-19 Vaccine Misinformation. Science Communication, 44(4), 387–417. https://doi.org/10.1177/10755470221111558
25. Jiang, L. C., Sun, M., Chu, T. H., & Chia, S. C. (2022). Inoculation works and health advocacy backfires: Building resistance to COVID-19 vaccine misinformation in a low political trust context. Frontiers in Psychology, 13, 976091.
26. Piltch-Loeb, R., Su, M., Hughes, B., Testa, M., Goldberg, B., Braddock, K., Miller-Idriss, C., Maturo, V., & Savoia, E. (2022). Testing the Efficacy of attitudinal inoculation videos to enhance COVID-19 vaccine acceptance: Quasi-experimental intervention trial. JMIR Public Health and Surveillance, 8(6), e34615.
27. Vandormael, A., Adam, M., Greuel, M., Gates, J., Favaretti, C., Hachaturyan, V., & Bärnighausen, T. (2021). The effect of a wordless, animated, social media video intervention on COVID-19 prevention: Online randomized controlled trial. JMIR Public Health and Surveillance, 7(7), e29060.
28. Agley, J., Xiao, Y., Thompson, E. E., Chen, X., & Golzarri-Arroyo, L. (2021). Intervening on trust in science to reduce belief in COVID-19 misinformation and increase COVID-19 preventive behavioral intentions: Randomized controlled trial. Journal of Medical Internet Research, 23(10), e32425.
29. Van Stekelenburg, A., Schaap, G., Veling, H., & Buijzen, M. (2021). Investigating and improving the accuracy of US citizens' beliefs about the COVID-19 pandemic: Longitudinal survey study. Journal of Medical Internet Research, 23(1), e24069.

30. Johnson, V., Butterfuss, R., Kim, J., Orcutt, E., Harsch, R., & Kendeou, P. (2022). The 'Fauci Effect': Reducing COVID-19 misconceptions and vaccine hesitancy using an authentic multimodal intervention. Contemporary Educational Psychology, 70, 102084.

31. Freeman, D., Loe, B. S., Yu, L.-M., Freeman, J., Chadwick, A., Vaccari, C., Shanyinde, M., Harris, V., Waite, F., & Rosebrock, L. (2021). Effects of different types of written vaccination information on COVID-19 vaccine hesitancy in the UK (OCEANS-III): A single-blind, parallel-group, randomised controlled trial. The Lancet Public Health, 6(6), e416–e427.

32. Bender, F. L., Rief, W., Brück, J., & Wilhelm, M. (2023). Effects of a video-based positive side-effect information framing: An online experiment. Health Psychology. https://psycnet.apa.org/record/2023-44789-001

33. Ma, J., Chen, Y., Zhu, H., & Gan, Y. (2023). Fighting COVID-19 misinformation through an online game based on the inoculation theory: Analyzing the mediating effects of perceived threat and persuasion knowledge. International Journal of Environmental Research and Public Health, 20(2), 980.

34. Fung, M. Y., Lee, Y. H., Lee, Y. T. A., Wong, M. L., Li, J. T. S., Nok Ng, E. E., & Lee, V. W. Y. (2022). Feasibility of a telephone-delivered educational intervention for knowledge transfer of COVID-19-related information to older adults in Hong Kong: A pre–post-pilot study. Pilot and Feasibility Studies, 8(1), 228. https://doi.org/10.1186/s40814-022-01169-y

35. Van Bavel JJ, Harris EA, Pärnamets P, Rathje S, Doell KC, Tucker JA. Political psychology in the digital (mis) information age: A model of news belief and sharing. Social Issues and Policy Review. 2021 Jan;15(1):84-113.

36. Gawronski, B. (2021). Partisan bias in the identification of fake news. Trends in Cognitive Sciences, 25(9), 723–724. https://doi.org/10.1016/j.tics.2021.05.001

37. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71

Support

This study took place as part of the PhD candidacy of Robert Dickinson at the University of Sussex, for which he is self-funded. No additional funding was utilised in this study. Non-financial support came from project supervisors Dominique Mackowski, Harm Van Marwijk, and Elizabeth Ford. Additionally, Katie Goddard performed the role of deselection replication as laid out in the methodology.

Competing interests

There are no competing interests to report.

Supporting information captions

[formatting requirements waived until Minor Revision decision received]

Appendices

1.1 Quality Assessment Table

Article	Cohort	Control or comparison group	Pre-post intervention data	Random assignment of participants to the intervention	Random selection of participants for assessment	Follow-up rate 80% or higher	Comparison groups equivalent on demographics	Comparison groups equivalent at baseline on outcome measures	Total score (out of 8)
Agley et al. 2021	Yes	Yes	Yes	Yes	Yes	N/A	Yes	Yes	7
Amazeen et al. 2022	Yes	Yes	Yes	Yes	Yes	N/A	Yes	N/A	6
Amin et al. 2021	Yes	No	Yes	N/A	no	Yes	N/A	N/A	3
Aslett et al. 2022	Yes	Yes	Yes	Yes	Yes	Yes	Yes	N/A	7
Basol et al. 2020	Yes	Yes	Yes	Yes	Yes	N/A	Yes	no	6
Bender et al. 2023	Yes	Yes	No	Yes	no	N/A	Yes	no	4
DeGarmo et al. 2022	Yes	Yes	No	Yes	Yes	N/A	Yes	Yes	6
Dias et al. 2020	Yes	Yes	Yes	Yes	Yes	N/A	Unknown	No	5
Freeman et al. 2021	Yes	Yes	No	Yes	Yes	N/A	Yes	Yes	6
Fung et al. 2022	Yes	No	Yes	Yes	no	Yes	N/A	N/A	4
Gavin et al. 2022	No	Yes	No	Yes	Yes	N/A	Yes	N/A	4
Iles et al. 2022	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	8
Jiang et al. 2022	Yes	Yes	No	Yes	no	Yes	Yes	Yes	7
Johnson et al. 2022	Yes	Yes	No	Yes	Yes	Yes	Yes	Unknown	6
Kreps et al. 2022	Yes	Yes	No	Yes	Yes	N/A	Yes	N/A	5
Ma et al. 2023	Yes	Yes	Yes	Yes	no	no	Unknown	Unknown	4
Maertens et al.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	8

2021									
Maragh-Bass et al. 2022	Yes	No	Yes	N/A	no	Yes	N/A	N/A	3
Pennycook et al. 2020	Yes	Yes	No	Yes	Yes	Yes	Yes	Unknown	7
Piltch-Loeb et al. 2022	Yes	Yes	Yes	No	Yes	N/A	Yes	No	5
Stekelenburg et al. 2021	Yes	Yes	Yes	Yes	Yes	Yes	Yes	N/A	7
Ugarte et al. 2023	Yes	Yes	No	Yes	Yes	Yes	Yes	N/A	6
Vandormael et al 2021	Yes	Yes	Yes	Yes	Yes	N/A	Yes	Yes	7
Veletsianos et al. 2022	Yes	No	No	N/A	N/A	N/A	N/A	N/A	1
Vijaykumar et al. 2021	Yes	Yes	No	Yes	Yes	N/A	Yes	No	5
Yousuf et al. 2021	Yes	Yes	Yes	Yes	No	No	Yes	No	5

646