

1 Effectiveness of digital health interventions
2 against COVID-19 misinformation: a
3 systematic realist review of intervention trials
4

5 Robert Dickinson^{1¶*}

6 Dominique Makowski²

7 Harm van Marwijk¹

8 Elizabeth Ford¹

9

10 ¹Department of Primary Care and Public Health, Brighton and Sussex Medical School, Brighton, United
11 Kingdom

12 ²Department of Psychology, University of Sussex, Brighton, United Kingdom

13 Abstract

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

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

35 **Author Summary**

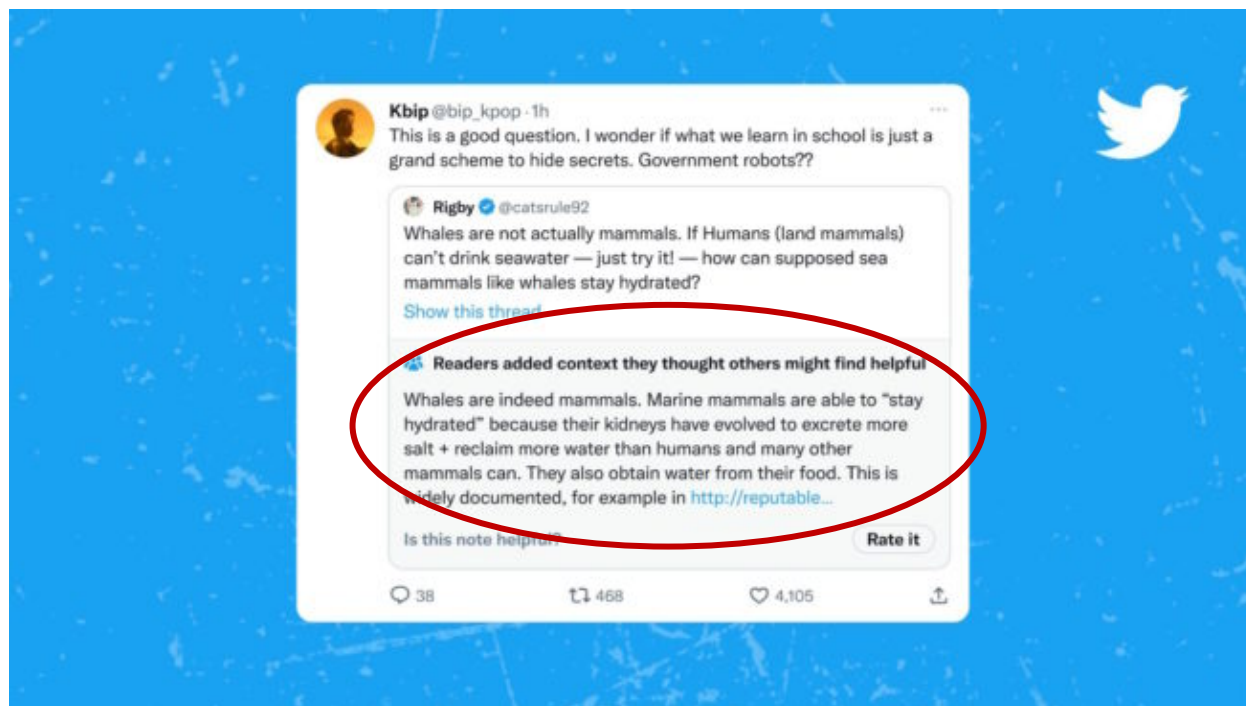
36 Misinformation is increasingly seen as a danger to public health and society at large. In the case of COVID-
37 19, it is associated with high levels of unnecessary death among the public. There have been many
38 interventions proposed to counter misinformation, yet little taking a meta-analytical perspective. These
39 interventions vary greatly and are not measured for effect in the same ways, making traditional comparisons
40 difficult. Instead, we categorised the interventions by type and assessed them by impact, scalability, durability,
41 and which groups of people and contexts in which they best work. With this information for each type of
42 intervention, policymakers can then make packages of multiple interventions that best work in their
43 circumstances. Although game-type interventions stood out from the rest, no one intervention seems capable
44 of effectively countering misinformation by itself. Many interventions were found to work differently on
45 different groups of people, which reaffirms suggestions by some authors that political ideology is relevant to
46 how people respond to these interventions. In future research there is a need to more deeply investigate the
47 role of politicisation in misinformation and interventions against it, as well as bringing in more theory to
48 understand how these interventions function.

49 **Introduction**

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

60 The pre-COVID misinformation intervention landscape was dominated by fact-checking. Fact-checking can
61 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



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

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

73 In the theoretical literature, discussion of misinformation interventions focused on inoculation, backfire
74 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?**

100 **RQ3: Under which circumstances are the interventions most effective?**

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

103 **Results**

104 **Study characteristics**

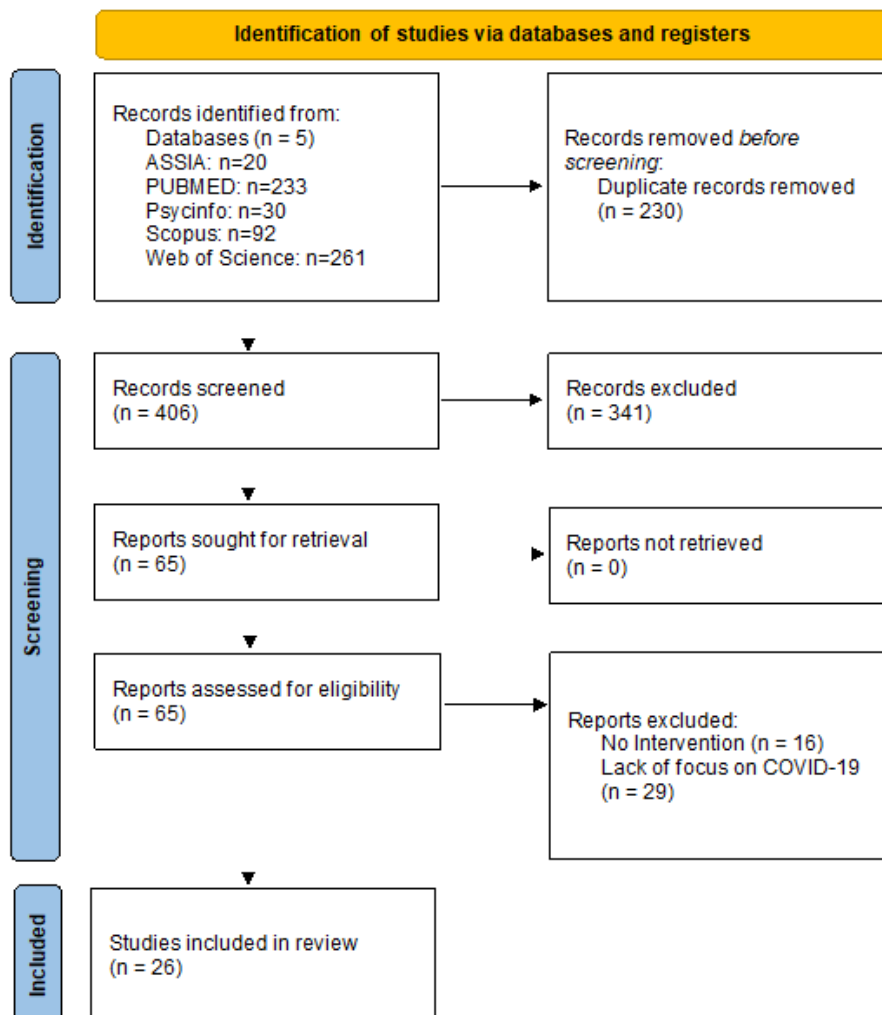
105 26 papers were found that met inclusion criteria, including 6 out of reviewing the bibliographies of the 20
106 studies found through the search strategy described above. 636 were initially resulting from the searches, with
107 230 duplicate results removed, 341 deselected by title, and 45 deselected by full-text review, resulting in 26
108 eligible papers (Figure 1). Papers were published between 2020 and 2023, with a variety of national, regional,
109 and international participant groups and study origination countries. The papers reviewed utilised participant
110 groups coming mainly from the USA through private research participant companies like MTurk, Lucid,
111 Prolific, Pollfish, and YouGov but also targeted audiences within the US like essential workers [10] and
112 'Latinx' communities [11]. Beyond the US, participant groups from Germany, the UK, Hong Kong, China,
113 Canada, the Netherlands, Brazil, Kyrgyzstan, India, and internationally were included in the reviewed studies.
114 These studies split into the intervention framework developed in the data extraction process as follows: 6
115 studies using Accuracy Nudges; 6 using education; 3 using Prebunking; 3 using Games; 3 using message
116 framing; 3 using Community Engagement; and 2 using Debunking (Table 1). Full details of the studies can be
117 seen in the appendices.

118 **Figure 1. PRISMA flow diagram for systematic reviews**

119 **Table 1. Study characteristics table**

120

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 significantly 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	randomized 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

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

130 **Intervention characteristics**

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

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

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

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

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

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

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

180 **Intervention effectiveness**

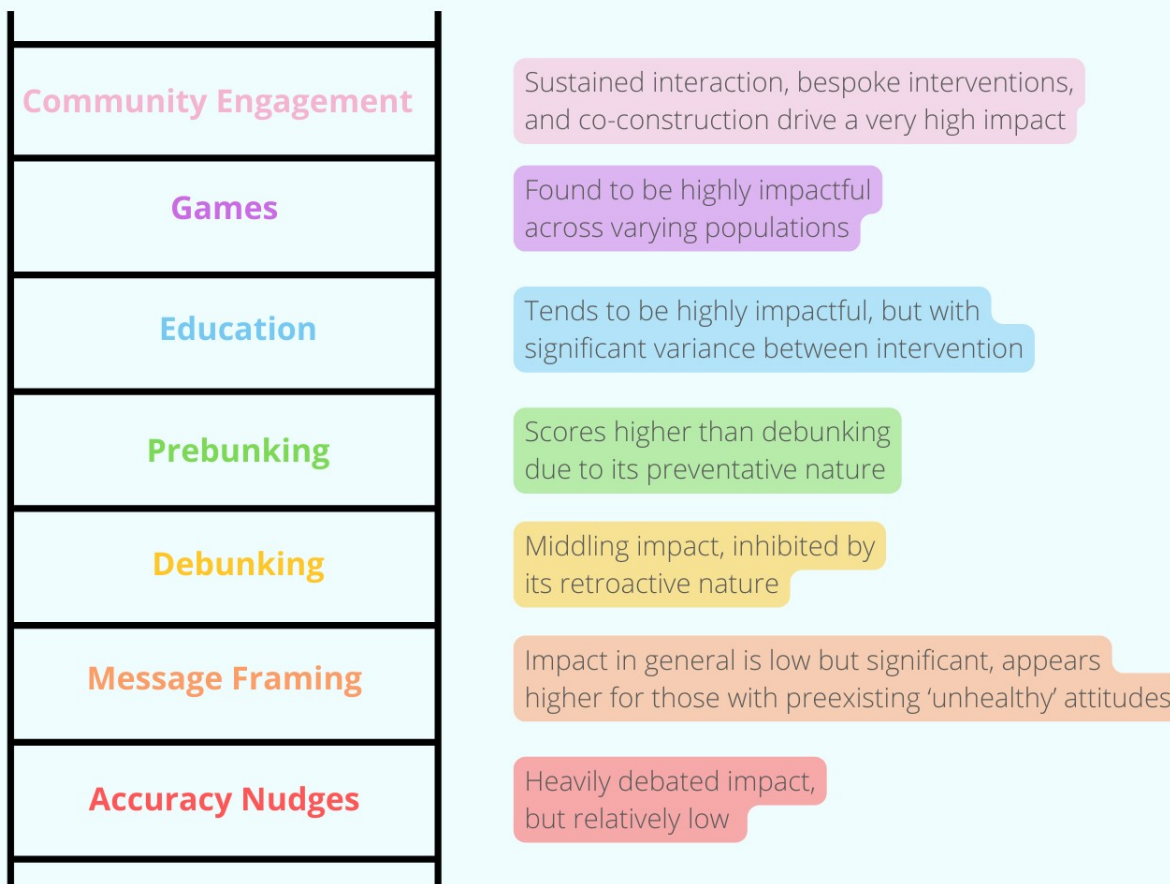
181 The two variables most central to answering which interventions work 'best' appear to be scalability and
182 impact. If impact is too low, the intervention might not actually engender sufficient behavioural change in
183 participants to combat the misinformation. Similarly, if an intervention cannot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

191 **Figure 2: Ladder of Impact on Participants**

192 **Figure 3: Ladder of Scalability**

193

Ladder of Impact on Participants



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

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

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

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

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

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

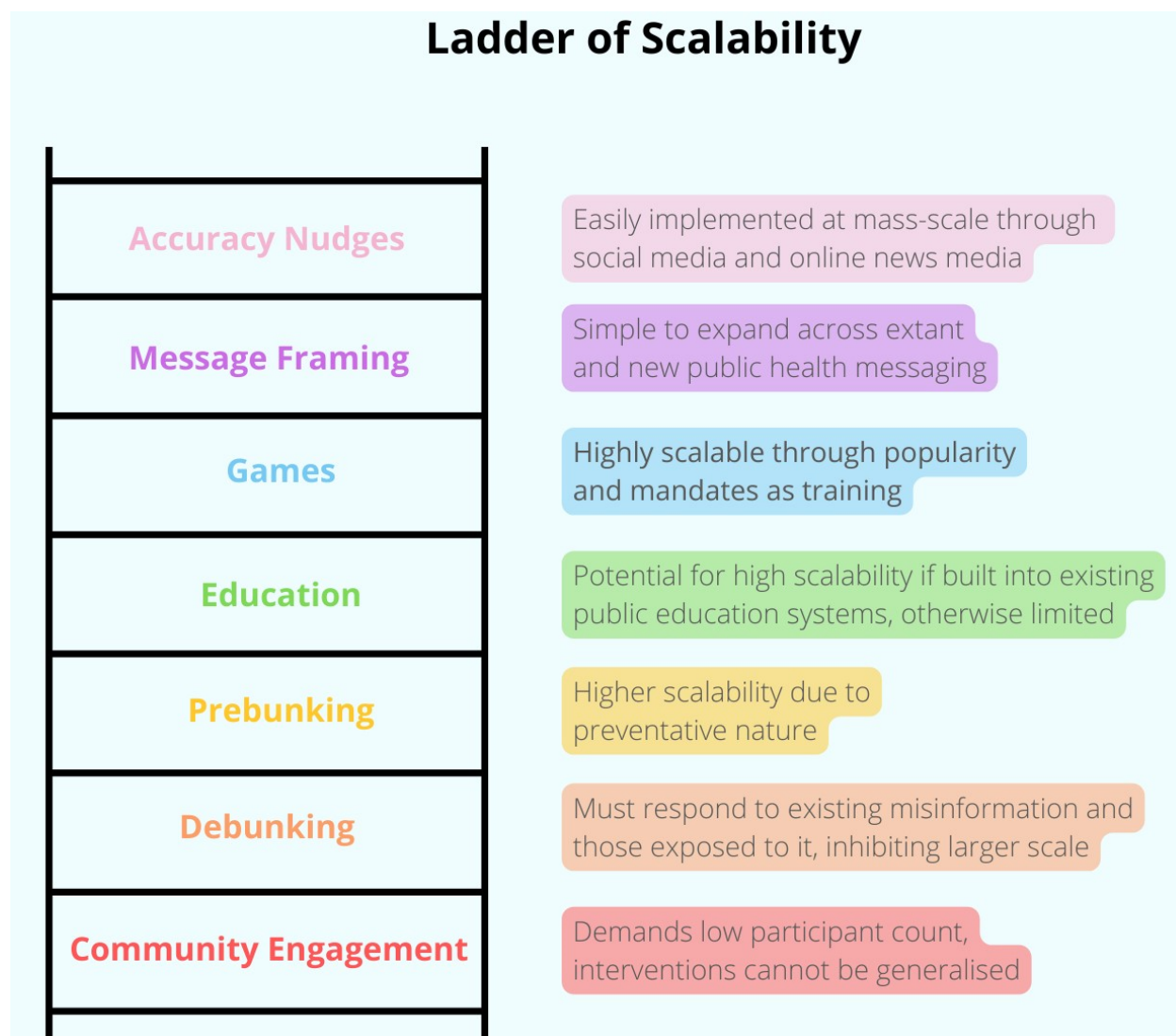
224 Games were consistently found to be highly impactful across the various populations who played them, with
225 high levels of durability and longevity compared to other intervention-types reviewed and significant impact
226 levels for all kinds of preexisting attitudes towards vaccination and COVID-19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



236

237 Community engagement is essentially impossible to scale upwards. It inherently requires small numbers of

238 participants and high levels of resource and time investment by those implementing the intervention. The

239 interventions themselves are then not even intended to be generalisable, but rather bespoke to and befitting

240 the contextual needs of the community involved. Community engagement can only effectively be done at a

241 small scale over long periods of time involving the building of trust with community, the proactive

242 engagement and co-creation of interventions and implementation strategies with the community itself, and

243 the implementation strategies themselves can take years to accomplish [10, 20].

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

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

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

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

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

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

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

270 Durability of effect

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

280 Special groups and circumstances

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

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

291 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	<ol style="list-style-type: none"> 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) 	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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) 	<ol style="list-style-type: none"> 1. Intervention only works on those with preexisting healthy attitudes 	High
Games	Youth, digitally literate people, employment mandates	<ol style="list-style-type: none"> 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) 	<ol style="list-style-type: none"> 1. Intervention works across the political spectrum 2. Intervention works well for general public 	High
Debunking	Reactively to widely believed	<ol style="list-style-type: none"> 1. UK and Brazil Whatsapp users (Vijaykumar et al. 2021) 	<ol style="list-style-type: none"> 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 (Agle 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)		

292

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

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

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

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

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

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

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

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

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

354 Discussion

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

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

366 **Politics and partisan bias**

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

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

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

399 **Limitations**

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

410 **Future Research Directions**

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

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

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

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

426 **Conclusions**

427 This review included 26 studies of interventions combatting COVID-19 misinformation. The interventions
428 reviewed varied widely in terms of scalability, resource intensity, impact on participants, the contexts within
429 which each best works, the people onto whom the interventions will have greatest effect, and research quality.

430 The tests performed in the included studies hold rich contributions toward better understanding how
431 misinformation functions, how veracity judgement occurs in individuals and communities, and which
432 interventions work best in which contexts and for whom. COVID-19 showed precisely how harmful and
433 deadly misinformation can be, and what a public health threat it can represent. In this fight against systemic
434 misinformation in our society, a final takeaway from this review is the need for acknowledgement of
435 misinformation as a societal and systemic issue that requires significant investment and time to resolve, if
436 resolution is possible.

437 **Materials and methods**

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

447 **Search strategy**

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

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

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

462 Eligibility

463 Trials or experimental studies were eligible if they were focused on reducing the spread of and vulnerability
464 to COVID-19 misinformation in their participants, and tested an intervention meant to combat COVID-19
465 misinformation. Studies were required to be in the English language and been published between 2020 and
466 2023 as searching before the COVID-19 pandemic began was unnecessary.

467 Quality assessment

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

476 Data extraction and analyses

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

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

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

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

504 Acknowledgments

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

508 References

- 509 1. Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis
510 of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, 28(11),
511 1531–1546. <https://doi.org/10.1177/0956797617714579>
- 512 2. Cook, J., Ecker, U., & Lewandowsky, S. (2015). Misinformation and How to Correct It. In R. A.
513 Scott & S. M. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences* (1st ed., pp. 1–
514 17). Wiley. <https://doi.org/10.1002/9781118900772.etrds0222>
- 515 3. Cohen, D. (2023, January 19). Twitter Extends Community Notes to Quote Tweets.
516 <https://www.adweek.com/media/twitter-extends-community-notes-to-quote-tweets/>
- 517 4. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19
518 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge
519 Intervention. *Psychological Science*, 31(7), 770–780. <https://doi.org/10.1177/0956797620939054>
- 520 5. Rathje, S., Roozenbeek, J., Traberg, C. S., Van Bavel, J. J., & van der Linden, S. (2022). Letter to the
521 editors of *Psychological Science*: Meta-analysis reveals that accuracy nudges have little to no effect
522 for US conservatives: regarding Pennycook et al.(2020).
523 <https://psyarxiv.com/945na/download?format=pdf>
- 524 6. Lees, J., McCarter, A., & Sarno, D. M. (2022). Twitter’s disputed tags may be ineffective at reducing
525 belief in fake news and only reduce intentions to share fake news among Democrats and
526 Independents. *Journal of Online Trust and Safety*, 1(3).
527 <https://www.tsjournal.org/index.php/jots/article/view/39>

- 528 7. Gawronski, B., Ng, N. L., & Luke, D. M. (2023). Truth sensitivity and partisan bias in responses to
529 misinformation. *Journal of Experimental Psychology: General*.
530 <https://psycnet.apa.org/record/2023-58364-001>
- 531 8. Gavin, L., McChesney, J., Tong, A., Sherlock, J., Foster, L., & Tomsa, S. (2022). Fighting the Spread
532 of COVID-19 Misinformation in Kyrgyzstan, India, and the United States: How Replicable Are
533 Accuracy Nudge Interventions? <https://assets.pubpub.org/d6zy4hsf/21663782255673.pdf>
- 534 9. Featherstone, J. D., & Zhang, J. (2020). Feeling angry: The effects of vaccine misinformation and
535 refutational messages on negative emotions and vaccination attitude. *Journal of Health
536 Communication, 25*(9), 692–702. <https://doi.org/10.1080/10810730.2020.1838671>
- 537 10. Ugarte, D. A., & Young, S. (2023). Effects of an Online Community Peer-support Intervention on
538 COVID-19 Vaccine Misinformation Among Essential Workers: Mixed-methods Analysis. *Western
539 Journal of Emergency Medicine, 24*(2), 264.
- 540 11. DeGarmo, D. S., De Anda, S., Cioffi, C. C., Tavalire, H. F., Searcy, J. A., Budd, E. L., McWhirter, E.
541 H., Mauricio, A. M., Halvorson, S., & Beck, E. A. (2022). Effectiveness of a COVID-19 testing
542 outreach intervention for Latinx communities: A cluster randomized trial. *JAMA Network Open,*
543 *5*(6), e2216796–e2216796.
- 544 12. Kennedy, C. E., Fonner, V. A., Armstrong, K. A., Denison, J. A., Yeh, P. T., O'Reilly, K. R., &
545 Sweat, M. D. (2019). The Evidence Project risk of bias tool: Assessing study rigor for both
546 randomized and non-randomized intervention studies. *Systematic Reviews, 8*(1), 3.
547 <https://doi.org/10.1186/s13643-018-0925-0>
- 548 13. Iles, I. A., Gaysynsky, A., & Sylvia Chou, W.-Y. (2022). Effects of Narrative Messages on Key
549 COVID-19 Protective Responses: Findings From a Randomized Online Experiment. *American
550 Journal of Health Promotion, 36*(6), 934–947. <https://doi.org/10.1177/08901171221075612>
- 551 14. Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of
552 inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental
553 Psychology: Applied, 27*(1), 1.

- 554 15. Veletsianos, G., Houlden, S., Hodson, J., Thompson, C. P., & Reid, D. (2022). An Evaluation of a
555 Microlearning Intervention to Limit COVID-19 Online Misinformation. *Journal of Formative*
556 *Design in Learning*, 6(1), 13–24. <https://doi.org/10.1007/s41686-022-00067-z>
- 557 16. Amin, Z., Ali, N. M., & Smeaton, A. F. (2021). Visual Selective Attention System to Intervene User
558 Attention in Sharing COVID-19 Misinformation (arXiv:2110.13489). arXiv.
559 <http://arxiv.org/abs/2110.13489>
- 560 17. Aslett, K., Guess, A. M., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). News credibility labels have
561 limited average effects on news diet quality and fail to reduce misperceptions. *Science Advances*,
562 8(18), eabl3844. <https://doi.org/10.1126/sciadv.abl3844>
- 563 18. Dias, N., Pennycook, G., & Rand, D. G. (2020). Emphasizing publishers does not effectively reduce
564 susceptibility to misinformation on social media.
565 [https://dspace.mit.edu/bitstream/handle/1721.1/144236/V2_researcharticle_publishers_jan29.pdf?](https://dspace.mit.edu/bitstream/handle/1721.1/144236/V2_researcharticle_publishers_jan29.pdf?sequence=2&isAllowed=y)
566 [sequence=2&isAllowed=y](https://dspace.mit.edu/bitstream/handle/1721.1/144236/V2_researcharticle_publishers_jan29.pdf?sequence=2&isAllowed=y)
- 567 19. Kreps, S. E., & Kriner, D. L. (2022). The COVID-19 infodemic and the efficacy of interventions
568 intended to reduce misinformation. *Public Opinion Quarterly*, 86(1), 162–175.
- 569 20. Maragh-Bass, A., Comello, M. L., Tolley, E. E., Stevens Jr, D., Wilson, J., Toval, C., Budhwani, H., &
570 Hightow-Weidman, L. (2022). Digital storytelling methods to empower young Black adults in
571 COVID-19 vaccination decision-making: Feasibility study and demonstration. *JMIR Formative*
572 *Research*, 6(9), e38070.
- 573 21. Vijaykumar, S., Jin, Y., Rogerson, D., Lu, X., Sharma, S., Maughan, A., Fadel, B., de Oliveira Costa,
574 M. S., Pagliari, C., & Morris, D. (2021). How shades of truth and age affect responses to COVID-19
575 (Mis) information: Randomized survey experiment among WhatsApp users in UK and Brazil.
576 *Humanities and Social Sciences Communications*, 8(1). [https://www.nature.com/articles/s41599-](https://www.nature.com/articles/s41599-021-00752-7)
577 [021-00752-7](https://www.nature.com/articles/s41599-021-00752-7)
- 578 22. Yousuf, H., van der Linden, S., Bredius, L., van Essen, G. T., Sweep, G., Preminger, Z., van Gorp,
579 E., Scherder, E., Narula, J., & Hofstra, L. (2021). A media intervention applying debunking versus

- 580 non-debunking content to combat vaccine misinformation in elderly in the Netherlands: A digital
581 randomised trial. *EClinicalMedicine*, 35.
582 [https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370\(21\)00161-9/fulltext](https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(21)00161-9/fulltext)
- 583 23. Basol, M., Roozenbeek, J., & Van der Linden, S. (2020). Good news about bad news: Gamified
584 inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, 3(1).
585 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6952868/>
- 586 24. Amazeen, M. A., Krishna, A., & Eschmann, R. (2022). Cutting the Bunk: Comparing the Solo and
587 Aggregate Effects of Prebunking and Debunking Covid-19 Vaccine Misinformation. *Science*
588 *Communication*, 44(4), 387–417. <https://doi.org/10.1177/10755470221111558>
- 589 25. Jiang, L. C., Sun, M., Chu, T. H., & Chia, S. C. (2022). Inoculation works and health advocacy
590 backfires: Building resistance to COVID-19 vaccine misinformation in a low political trust context.
591 *Frontiers in Psychology*, 13, 976091.
- 592 26. Piltch-Loeb, R., Su, M., Hughes, B., Testa, M., Goldberg, B., Braddock, K., Miller-Idriss, C., Maturo,
593 V., & Savoia, E. (2022). Testing the Efficacy of attitudinal inoculation videos to enhance COVID-19
594 vaccine acceptance: Quasi-experimental intervention trial. *JMIR Public Health and Surveillance*, 8(6),
595 e34615.
- 596 27. Vandormael, A., Adam, M., Greuel, M., Gates, J., Favaretti, C., Hachaturyan, V., & Bärnighausen, T.
597 (2021). The effect of a wordless, animated, social media video intervention on COVID-19
598 prevention: Online randomized controlled trial. *JMIR Public Health and Surveillance*, 7(7), e29060.
- 599 28. Agle, J., Xiao, Y., Thompson, E. E., Chen, X., & Golzarri-Arroyo, L. (2021). Intervening on trust in
600 science to reduce belief in COVID-19 misinformation and increase COVID-19 preventive
601 behavioral intentions: Randomized controlled trial. *Journal of Medical Internet Research*, 23(10),
602 e32425.
- 603 29. Van Stekelenburg, A., Schaap, G., Veling, H., & Buijzen, M. (2021). Investigating and improving the
604 accuracy of US citizens' beliefs about the COVID-19 pandemic: Longitudinal survey study. *Journal*
605 *of Medical Internet Research*, 23(1), e24069.

- 606 30. Johnson, V., Butterfuss, R., Kim, J., Orcutt, E., Harsch, R., & Kendeou, P. (2022). The ‘Fauci
607 Effect’: Reducing COVID-19 misconceptions and vaccine hesitancy using an authentic multimodal
608 intervention. *Contemporary Educational Psychology*, 70, 102084.
- 609 31. Freeman, D., Loe, B. S., Yu, L.-M., Freeman, J., Chadwick, A., Vaccari, C., Shanyinde, M., Harris, V.,
610 Waite, F., & Rosebrock, L. (2021). Effects of different types of written vaccination information on
611 COVID-19 vaccine hesitancy in the UK (OCEANS-III): A single-blind, parallel-group, randomised
612 controlled trial. *The Lancet Public Health*, 6(6), e416–e427.
- 613 32. Bender, F. L., Rief, W., Brück, J., & Wilhelm, M. (2023). Effects of a video-based positive side-effect
614 information framing: An online experiment. *Health Psychology*.
615 <https://psycnet.apa.org/record/2023-44789-001>
- 616 33. Ma, J., Chen, Y., Zhu, H., & Gan, Y. (2023). Fighting COVID-19 misinformation through an online
617 game based on the inoculation theory: Analyzing the mediating effects of perceived threat and
618 persuasion knowledge. *International Journal of Environmental Research and Public Health*, 20(2),
619 980.
- 620 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.
621 (2022). Feasibility of a telephone-delivered educational intervention for knowledge transfer of
622 COVID-19-related information to older adults in Hong Kong: A pre–post-pilot study. *Pilot and
623 Feasibility Studies*, 8(1), 228. <https://doi.org/10.1186/s40814-022-01169-y>
- 624 35. Van Bavel JJ, Harris EA, Pärnamets P, Rathje S, Doell KC, Tucker JA. Political psychology in the
625 digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review*.
626 2021 Jan;15(1):84-113.
- 627 36. Gawronski, B. (2021). Partisan bias in the identification of fake news. *Trends in Cognitive Sciences*,
628 25(9), 723–724. <https://doi.org/10.1016/j.tics.2021.05.001>
- 629 37. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA
630 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 2021;372:n71. doi:
631 10.1136/bmj.n71

632 **Support**

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

637

638 **Competing interests**

639 There are no competing interests to report.

640

641 **Supporting information captions**

642 [formatting requirements waived until Minor Revision decision received]

643 **Appendices**

644

645 1.1 Quality Assessment Table

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