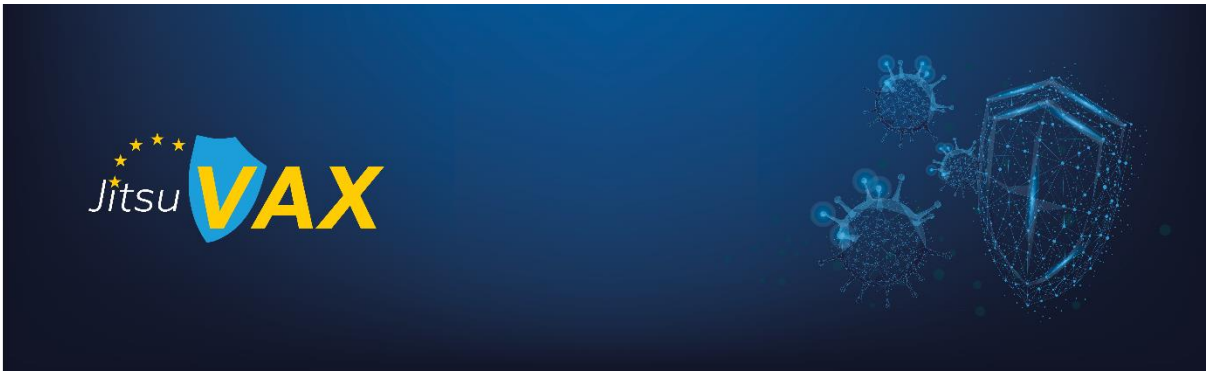


JITSUVAX:
Jiu-Jitsu with Misinformation in the Age of Covid

Reports on Task WP2.2 (inoculation and diffusion)

June 2022

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 964728 (JITSUVAX)

Co-funded by the Horizon 2020 programme of the European Union

JITSUVAX Deliverable 2.2

Reports on Task WP2.2 (inoculation and diffusion)

Project title:	JITSUVAX: Jiu-Jitsu with Misinformation in the Age of Covid
Grant agreement:	964728
Duration:	April 2021-March 2025
Website:	https://sks.to/jitsuvax
Coordinator:	Stephan Lewandowsky
Deliverable number:	2.2
Deliverable Title:	Reports on Task WP2.2 (inoculation and diffusion)
Dissemination level:	Public
Version:	1
Authors:	Rakoen Maertens, Jon Roozenbeek, Sander van der Linden
Reviewed by:	Stephan Lewandowsky
Contacts:	rm938@cam.ac.uk, jitsuvax@bristol.ac.uk

Consortium: **University of Bristol**, Beacon House Queens Road, Bristol, BS8 1QU, UK
Universität Erfurt, Nordhauser Strasse 63, Erfurt 99089, Germany
The Chancellor Masters and Scholars of the University of Cambridge, Trinity Lane, The Old Schools, Cambridge, CB2 1TN, UK
Turun yliopisto, Yliopistonmaki, Turku 20014, Finland
Observatoire Regional de la Sante,27 Boulevard Jean Moulin, Marseille 13005, France
Universidade de Coimbra, Paço das Escolas, Coimbra 3001 451, Portugal.

Contents

Summary

Psychological "inoculation" has been widely implemented as a method to counter the influence of misinformation by preemptively exposing people to the misleading information they could encounter (Lewandowsky & van der Linden, 2021). This is done in a controlled setting that teaches participants about the flaws in misinformation and warns them to not get influenced by them. Two gamified inoculation interventions (*BadVaxx* and *VaxBN*) were also developed as part of the JITSUVAX project (WP2.2). While inoculation interventions have been widely implemented and meta-analyses show that they are successful at increasing resistance to misinformation, a challenge remains to spread this "psychological vaccine" (inoculation interventions) in a way that matches the spread of the "psychological virus" (misinformation).

A promising and zero-cost avenue to explore is whether the effects of inoculation interventions spread beyond those who have been inoculated. One such approach is to look at whether former inoculation intervention participants are talking about what they have learned and thereby protecting others in their network (friends, family, peers). A first step towards exploring this was taken in JITSUVAX Deliverable 3.3, where post-inoculation talk was explored in the context of discussions on online forums after a session of the *Bad News* intervention (a gamified inoculation intervention related to *Bad Vaxx*).

In the present project we take this concept one step further and experimentally expose people to "diffusion messages": messages written by participants who went through an inoculation intervention specifically crafted to protect others. We explore how exposing people who never went through an inoculation intervention are affected by reading such diffusion messages and track people over time within a 30-day time frame. In addition, we explore this further by looking at the inoculation effect's long-term effectiveness curve and the potential benefits of an inoculation booster intervention, to put the effects into context. Finally, we look at another level beyond that: 2nd degree diffusion messages (diffusion messages written by those who have read a diffusion message), and whether exposing people to a 2nd degree diffusion message has any beneficial effects.

In this study, we find that Bad Vaxx works as robustly with effects lasting up to 10 days without any booster intervention. When participants are asked to create diffusion messages, they in general create diffusion messages that incorporate some elements of the inoculation intervention, reflecting that they are able to convey some of the materials they have learned in the game. When exposing a control group to the 1st degree diffusion messages, we find a slight descriptive increase in misinformation discernment performance, although the effect is not significant and negligible compared to an actual inoculation intervention. When asking those participants to create a 2nd degree diffusion message, they are able to create a generic message that could be seen as a media literacy tip, but does not have much resemblance to the original inoculation intervention. In contrast, when a booster intervention is applied (a second inoculation), the intervention is meaningfully boosted, with the intervention lasting up to at least 4 weeks. In summary, this indicates that there is potential for gamified inoculation interventions to spread and protect others, but this potential is limited to direct peers of the participant of the original intervention, and the effects are – if present– limited compared to the benefit of engaging with an inoculation intervention directly.

Scope and purpose of this document

This document reports on a study conducted to evaluate whether gamified inoculation interventions such as *Bad Vaxx* have the ability to protect others through diffusion messages ("post-inoculation

talk"), i.e., talking about the content learned in the inoculation intervention to help protect peers. This document lays out the background, methodology, results, and other findings of this study.

Project overview

Vaccine hesitancy—the delay or refusal of vaccination without medical indication—has been cited as a serious threat to global health by the World Health Organization (WHO), attributing it to misinformation on the internet. The WHO has also identified Health Care Professionals (HCPs) as the most trusted influencers of vaccination decisions.

JITSUVAX will leverage those insights to turn toxic misinformation into a potential asset based on two premises:

1. The best way to acquire knowledge and to combat misperceptions is by employing misinformation itself, either in weakened doses as a cognitive "vaccine", or through thorough analysis of misinformation during "refutational learning".
2. HCPs form the critical link between vaccination policies and vaccine uptake.

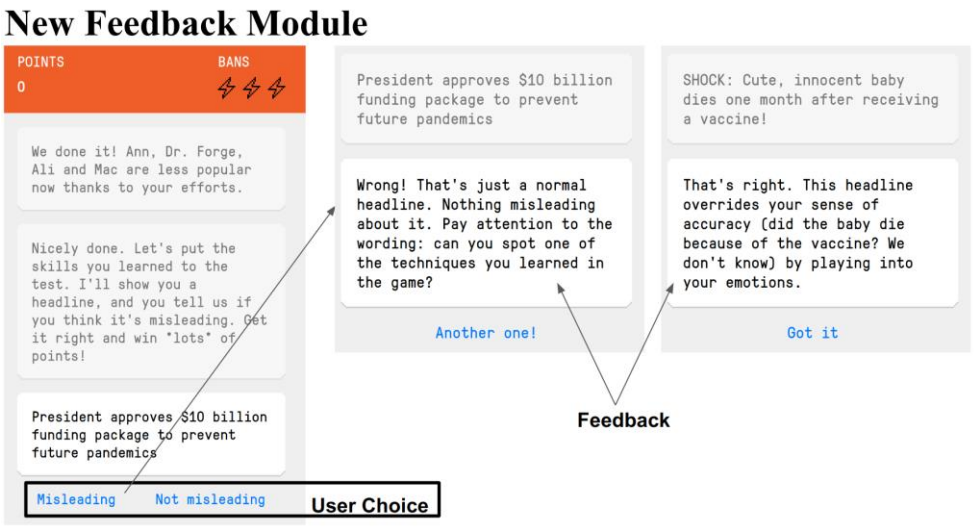
The principal objective of JITSUVAX is to leverage misinformation about vaccinations into an opportunity by training HCPs through inoculation and refutational learning, thereby neutralizing misinformation among HCPs and enabling them to communicate more effectively with patients. We will disseminate and leverage our new knowledge for global impact through the team's contacts and previous collaborations with WHO and UNICEF.

Background

Active inoculation through gamification

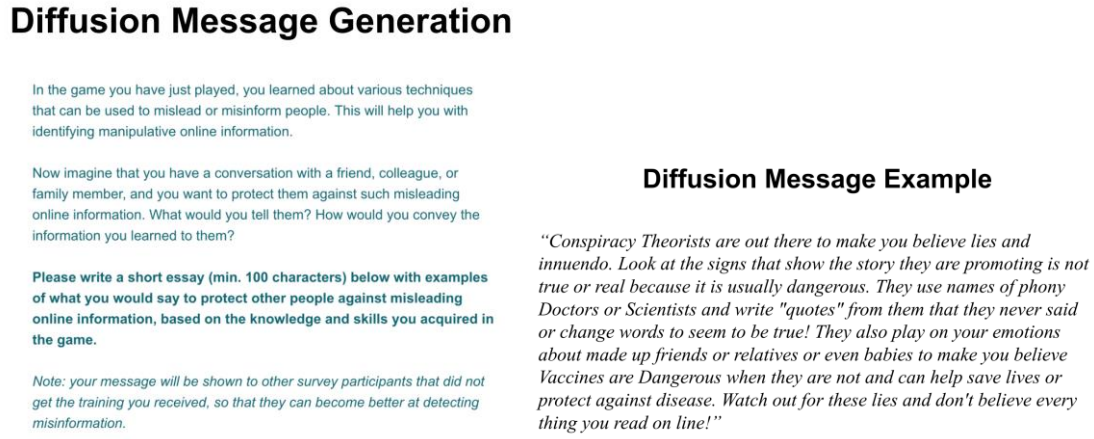
This project uses a gamified form of inoculation. In a 15-minute intervention called *Bad Vaxx* participants learn how misinformation is formed and spread on social media, and specifically, participants learn about four techniques of manipulation: (1) emotional storytelling, (2) pseudoscience and fake expertise, (3) naturalistic fallacies, and (4) conspiracy theories. The intervention was developed as part of the JITSUVAX project (WP2.2). For more information about how inoculation games are developed and look like, please refer to JITSUVAX Deliverable 3.3 (https://jitsuvax.github.io/files/D3.3%20Postinoculation%20talk.pdf).

Specifically for this study we use the "good version" of the Bad Vaxx game in which you are instructed to *detect* misinformation spread by others. In addition, we have added a new "feedback module" based on the latest insights into improving both the longevity and the discernment effectiveness of inoculation interventions (Capewell et al., 2023; Leder et al., 2023). The feedback module adds a short exercise at the end of the intervention where people have to indicate whether or not a headline is misleading, with real-time feedback. An example stimulus of a misleading and a neutral headline within the feedback module and the feedback for a correct or wrong response can be found in the figure below.



Post-inoculation talk and diffusion messages

While various studies have shown that inoculation is an effective way to boost people's ability to discern trustworthy messages and headlines from manipulative or misleading ones (Lu et al., 2023), challenges remain in the field of inoculating *enough* people to make society as a whole resilient to misinformation. Research has been done to explore whether people talk about the inoculation intervention's content to other people after they have been inoculated, and thereby also rehearse the information learned and stay motivated to protect themselves (Dillingham & Ivanov, 2016; Ivanov et al., 2012). In the JITSUVAX D3.3 report the presence of post-inoculation talk was already explored via the analysis of Reddit post data, and it was found that post-inoculation talk (incl. counter-arguing) was more prevalent after the inoculation intervention was shared, albeit with very small effects ($p < .001$, $d = \sim 0.035$). In this project we go one step further and explore whether we can experimentally expose people to post-inoculation talk.



The present study

In this experiment we investigated the long-term effectiveness of feedback-enhanced gamified inoculation interventions, and whether those who were inoculated are then able to protect others by telling them about what they learned (i.e., the spread of inoculation effects beyond the inoculated using "diffusion messages"), up to 2nd degree indirect inoculation. To do this, we investigated the effectiveness of the online game "Bad Vaxx" in a diffusion paradigm, where participants who played Bad Vaxx were asked to write a message to inoculate peers who did not receive the intervention. "1st degree indirect inoculation" was defined as people inoculated via

messages (1st degree diffusion messages) from people that received the full inoculation intervention, while "2nd degree indirect inoculation" was defined as people inoculated through messages generated by those who only got inoculated through 1st degree indirect inoculation (2nd degree diffusion messages).

Our hypotheses are based on a recent paper by Maertens et al. (2023) examined the long-term effectiveness of inoculation interventions in great detail, including the effects of booster interventions. They served as an inspiration for both our research design and as a basis for most of our hypotheses, which can be found below:

Main Effect Hypothesis

- **[H1]** Playing a short online feedback-enhanced game about vaccine misinformation improves people's ability to discern manipulative social media content about vaccinations 0–10 days[+] after the intervention.

Decay Hypotheses

- **[H2]** The inoculation effect remains significant 11–20 days[+] after the intervention [H2a], but is no longer significant 21–30 days[+] after the intervention [H2b].

- **[H3]** The inoculation effect decays quickly immediately after the intervention and slows down over time, approximating an exponential decay curve, reflected in an effect decay between 0–10 days[+] and 11–20 days[+] that is larger than the difference between 11–20 days[+] and 21–30 days[+]

Booster Hypothesis

- **[H4]** The inoculation effect remains significant 21–30 days[+] after the intervention if a booster intervention was administered 9–11 days after the intervention.

Diffusion Hypotheses

- **[H5]** The 1st degree indirect inoculation effect (via diffusion messages) is significant at 11–20 days[+] after the intervention [H5a], but no longer at 21–30 days[+] after the intervention [H5b].

- **[H6]** The 2nd degree indirect inoculation effect (via diffusion messages) is significant at 21–30 days[+] after the intervention.

[+] Based on the average effect across these days.

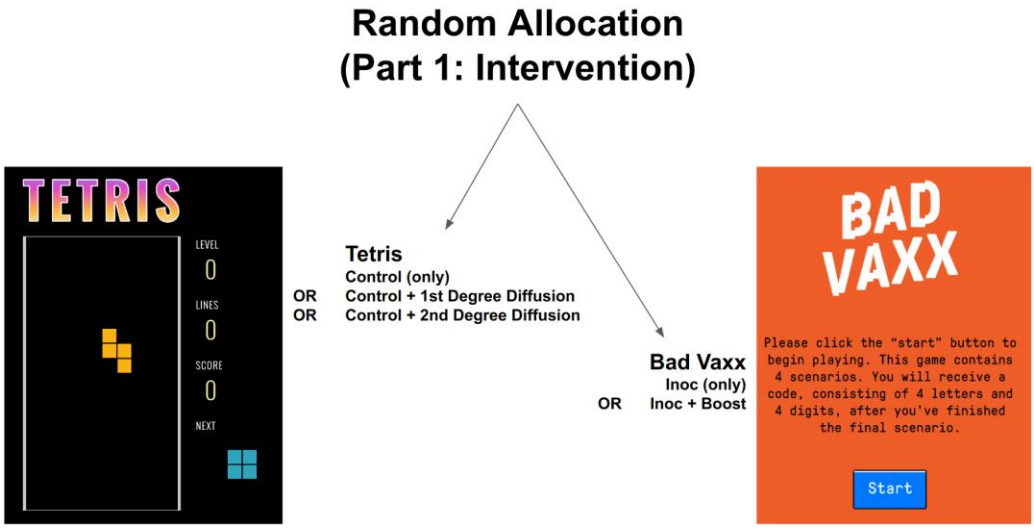
Methodology

Sample and Procedure

N = 8,525 (*n* = 1,705 per group) participants were recruited for this experiment. Based on a power analysis with effect size *d* = .20 (based on the meta-analytic effect size for manipulativeness discernment and the effect size for technique recognition in Appel et al., 2023, as well as the threshold for a small effect as the smallest effect size of interest), 55% participant attrition over time, power = .95, and alpha = .05, a minimum sample of *N* = 7,235 (*n* = 1,447 per group) is needed. An extra ~15% participants was recruited on top of this as there may be some unexpected participant drop-outs or participants not passing the quality checks.

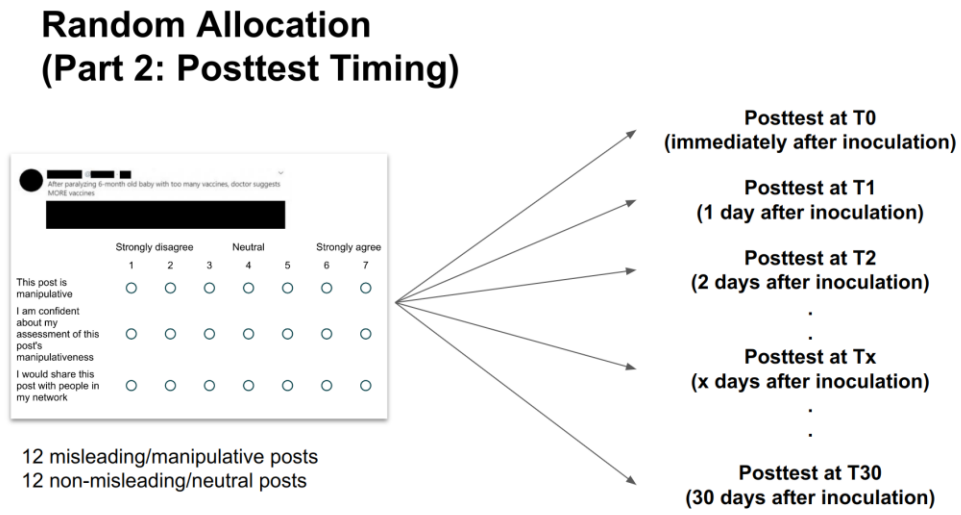
Participants were recruited from a US panel with balanced soft quota for age and gender recruited by the market research group Bilendi & respondi, after which they were invited to a Qualtrics survey that started with an informed consent. Then participants were allocated to one of 5 intervention

conditions: a control (Tetris) condition, 1st degree diffusion condition (Tetris and exposure to a 1st degree indirect inoculation message generated by inoculation group), a 2nd degree diffusion condition (Tetris and exposure to a 2nd degree indirect inoculation message generated by the 1st Degree Diffusion group), an inoculation condition (Bad Vaxx Game – good version with feedback module), or a inoculation boost condition (Bad Vaxx Game twice). The three conditions with Tetris all had the same T0 survey, as the 1st degree diffusion message was only presented at T10, and the 2nd degree diffusion message was only presented at T20, both to allow time for the participants from the other groups to generated the diffusion messages. Participants in the two inoculation conditions also had the same T0 survey, as the booster intervention was administered at T10.



Participants who participated in the first inoculation intervention were asked to write a short essay (called a "diffusion message") for participants who have not participated in the Bad Vaxx game (at T0), with the goal to protect them against misleading online information. These messages were a minimum of 100 characters (enforced) with no maximum. These messages were then shown to people in the 1st degree diffusion group at T9–T11, who then in turn created a similar message to protect the 2nd degree diffusion group at T19–T21. The messages shown to the participants were each time a random message from the database of generated diffusion messages. This database only contained quality-checked messages based on a quality check by authors RM and JR (a simple binary "relevant"/"irrelevant" categorisation was used and discussed until an inter-rater agreement of at least 90% is reached) for the 1st degree diffusion messages, and by RM only for the 2nd degree diffusion messages. Each participant saw only one diffusion message.

After the intervention, participants were invited to a possttest. This possttest consisted of an item rating task which enabled us to calculate the main outcome variable of interest (**"manipulativeness discernment")**. Participants were presented with a total of 12 fictitious social media posts. Each post was randomly either a post that contained a manipulation technique commonly used in vaccine misinformation (e.g. the use of fake experts) or its matched control post, similar in length and content, but without using manipulation. Participants were asked three questions for each post, all on a 1-7 scale (1 being "strongly disagree" and 7 being "strongly agree"): 1) this post is manipulative, 2) I am confident about my assessment of this post's manipulativeness, 3) I would share this post with people in my network. Using the manipulativeness ratings, we calculated the manipulativeness discernment score by subtracting the average scores of neutral posts from the average scores for the manipulative posts.



The timing of this posttest was randomised, and could be immediately on intervention day (T0), one day after the intervention (T1), two days after (T2), and so on until 30 days after the intervention (T30). Every participant only participated in one single posttest in order to eliminate any repeated testing confounds. This procedure led to a final set of five groups split over 31 potential posttest times, resulting in the design reflected in the list below (all groups are independent/separate groups, with the sample split equally 1/5):

- Group 1 (Control) – T0: Control | Tx^{++}: Posttest
- Group 2 (Diffusion1) – T0: Control | T10: Diffusion1 | Tx^{++}: Posttest
- Group 3 (Diffusion2) – T0: Control | T20: Diffusion2 | Tx^{++}: Posttest
- Group 4 (Inoc) – T0: Inoc | Tx^{++}: Posttest
- Group 5 (InocBoost) – T0: Inoc | T10: Boost | Tx^{++}: Posttest

$^{++}$ Tx is a random date in the range T0–T30. Every participant participated in only 1 posttest session.

Finally, for data analyses, cases were excluded if one or more of the exclusion conditions were met, which led to a final sample size of 3,805 (~761 participants per condition). The exclusion criteria were:

1) Participant participates in any of the surveys multiple times
   (i.e., all second+ entries of the same participant for each survey will be removed)
2) Participant does not accept the informed consent
3) Participant fails the manipulation check
4) Participant enters the wrong password after completing the Bad Vaxx intervention
5) Participant fails the attention check
6) Participant does not complete the entire survey (i.e., only complete cases will be accepted)
7) Participant does not participate in the booster or diffusion session within 3 days after invitation (i.e., within 9–11 or 19–21 days after T0)

Method of analysis

To test H1–H6 we ran a one-way ANOVA with discernment as the dependent variable and the intervention as the independent variable, using the T0–T10^{a} data for the H1 test, T11–T20^{a} for H2a, T21–T30^{a} for H2b, T0–T30^{a} for H3, T21–T30^{a} for H4, T11–T20^{a} data for H5, and T21–T30^{a} for H6. We then evaluated the below contrasts – all tests were two-sided tests – with a Tukey's HSD (honest significant difference) correction. We also planned some additional exploratory analyses, such as the plotting of a smooth decay curve of the inoculation effect and natural language processing to learn

more about the diffusion messages. A list of tests for each specific hypothesis can be found below (with significance in this report referring to a corrected *p*-value lower than $p = .05$):

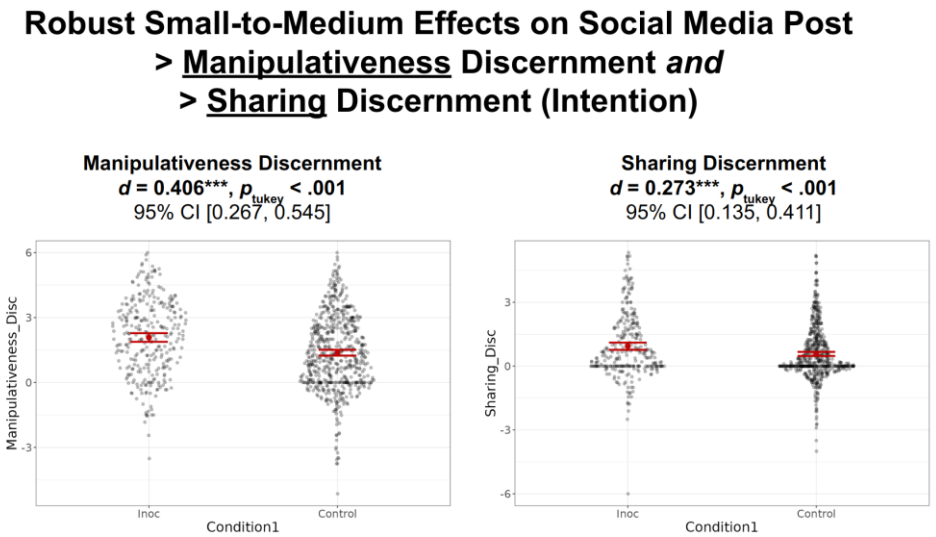
- **[H1]** Inoc (vs Control) leads to significant improvement in manipulativeness discernment when tested at T0–T10[a].
- **[H2a]** Inoc (vs Control) leads to significant improvement in manipulativeness discernment when tested at T11–T20[a].
- **[H2b]** Inoc (vs Control) does not lead to significant improvement in manipulativeness discernment when tested at T21–T30[a].
- **[H3]** The change in the Inoc (vs Control) effect between T0–T10[a] and T11–T20[a] in manipulativeness discernment is significantly larger compared to the change between T11–T20[a] and T21–T30[a].
- **[H4]** InocBoost (vs Control) effect on discernment is significant when tested at T21–T30[a].
- **[H5a]** Diffusion1 (vs Control) effect on discernment is significant when tested at T11–T20[a].
- **[H5b]** Diffusion1 (vs Control) effect on discernment is significant when tested at T21–T30[a].
- **[H6]** Diffusion2 (vs Control) effect on discernment is significant when tested at T21–T30[a].

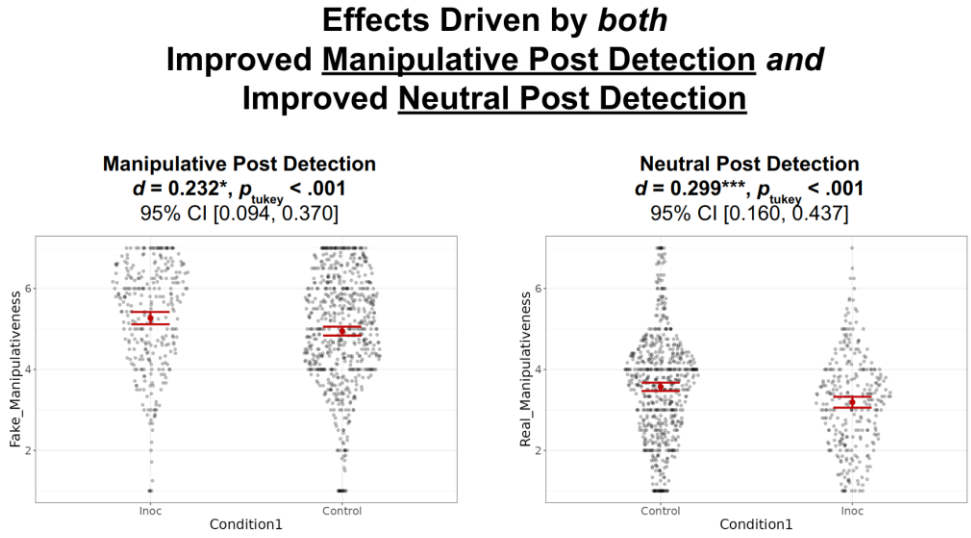
[a] Based on the average effect across these days.

Results

Overall Effectiveness

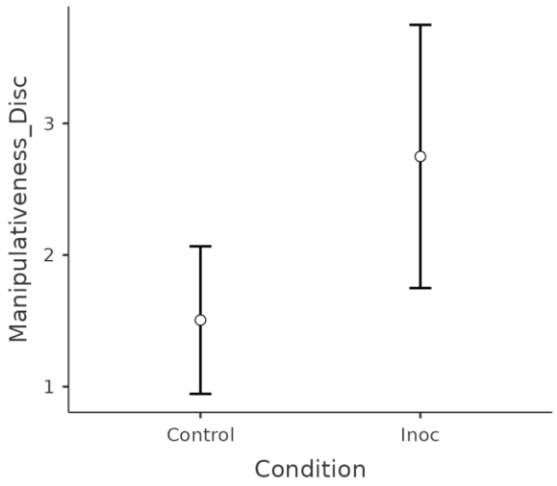
When taking into account the full dataset (across time points), robust small-to-medium effects are found for the discernment of manipulative from non-manipulative social media posts, both for manipulativeness ratings ($d = 0.406$***) and for sharing intentions ($d = 0.273$***; i.e., participants indicate they are more willing to share neutral social media posts than misleading ones). When looking into whether these effects are driven more by the detection of manipulative posts or by the detection of neutral posts, we find evidence for both: a boost in detecting manipulative content ($d = 0.406$***), and a boost in trusting non-manipulative content ($d = 0.273$***).





1-Week Effectiveness

The first analysis looks at the first 10 days after the inoculation intervention. On average, in this period, we find a strong baseline inoculation effect for the discernment of manipulativeness of stimuli with a medium-to-large effect size of Cohen's d = 0.756, 95% CI [0.040, 1.471], $t(44)$ = 2.186, p_{tukey} = .034 (**evidence for H1**). See the table figure below for an analysis.



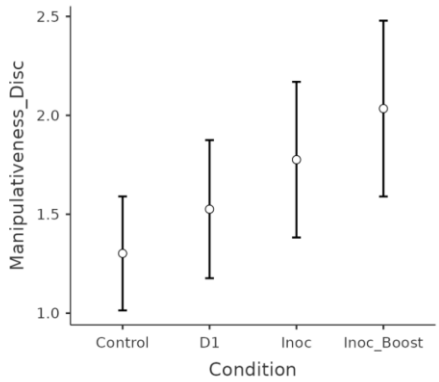
2-Week Effectiveness and First Degree Diffusion

Analysing the data 11–20 days after the inoculation intervention, including in the group that was only exposed to diffusion messages ("D1" for first degree diffusion), we find that the inoculation effect has decayed and is no longer significant (**evidence against H2**). The first degree diffusion message descriptively improved discernment skills after exposure – note that participants in the D1 condition were exposed to the diffusion message at T10 and thus the T11–T20 test for D1 is the equivalent of the T0–T10 test for the inoculation condition – but the effect was not significant (d = 0.130, p_{tukey} = .766, **evidence against H5**). This is potentially the case because the sample size was not large enough to detect effects smaller than d = 0.200. Only the "inoculation booster" condition, where participants went through the inoculation intervention one more time (also at T10), was significant with a small-to-medium effect size of Cohen's d = 0.425 (p_{tukey} = .035).

ANCOVA With T11–T20 Data

Comparison			Mean Difference	*SE*	*df*	*t*	p_{tukey}	Cohen's *d*	95% Confidence Interval	
Condition		**Condition**							*LL*	*UL*
Control	-	D1	0.224	0.230	360.000	0.972	.766	0.130	-0.133	0.393
	-	Inoc	0.474	0.248	360.000	1.911	.225	0.275	-0.009	0.559
	-	**Inoc_Boost**	**0.732**	**0.269**	**360.000**	**2.718**	**.035**	**0.425**	**0.116**	**0.735**
D1	-	Inoc	0.250	0.267	360.000	0.936	.786	0.145	-0.160	0.451
	-	Inoc_Boost	0.508	0.287	360.000	1.769	.290	0.295	-0.034	0.624
Inoc	-	Inoc_Boost	0.258	0.302	360.000	0.855	.828	0.150	-0.195	0.495

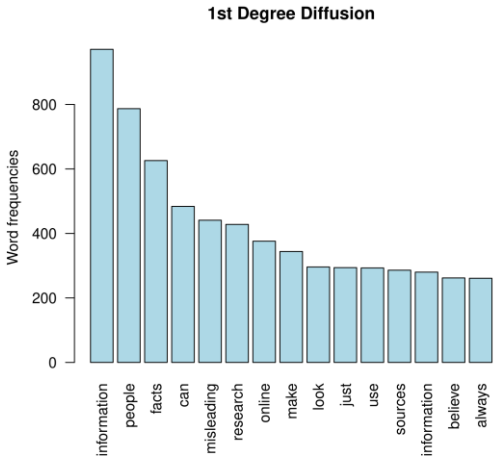
Only Effect Left at T11–T20 is for Inoculation Booster Condition, No Significant Effect for Diffusion Message Exposure



Manipulativeness Discernment

1st Degree Diffusion (D1)
d = 0.130, p_{tukey} = .766
95% CI [-0.133, 0.393]

Inoculation Booster (Inoc_Boost)
***d* = 0.425*, p_{tukey} = .035**
95% CI [0.116, 0.735]



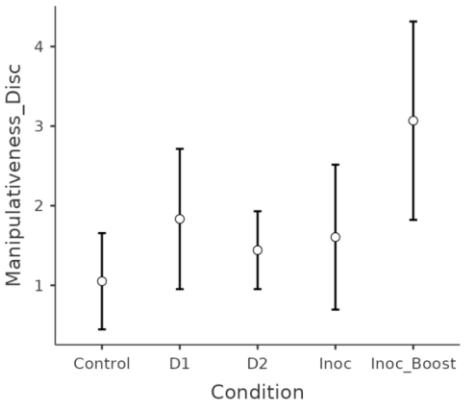
4-Week Effectiveness and Second Degree Diffusion

Finally, we looked at the data for the period of 21–30 days after the intervention. We found that none of the groups found any significant effects compared to the control group except for the inoculation booster group, which showed a very large effect size of Cohen's *d* = 1.133 (p_{tukey} = .037, **evidence for H4**). The second degree diffusion messages did not seem to make an impact on people's skills to discern manipulative stimuli from neutral stimuli (**evidence against H6**).

ANCOVA With T21–T30 Data

Comparison									95% Confidence Interval	
Condition		Condition	Mean Difference	*SE*	*df*	*t*	p_{tukey}	Cohen's *d*	*LL*	*UL*
Control	-	D1	0.782	0.540	120.000	1.449	.597	0.439	-0.163	1.042
	-	D2	0.390	0.393	120.000	0.994	.857	0.219	-0.218	0.657
	-	Inoc	0.554	0.552	120.000	1.004	.853	0.311	-0.304	0.926
	-	**Inoc_Boost**	**2.017**	**0.699**	**120.000**	**2.884**	**.037**	**1.133**	**0.342**	**1.925**
D1	-	D2	-0.392	0.509	120.000	-0.770	.939	-0.220	-0.787	0.347
	-	Inoc	-0.228	0.640	120.000	-0.356	.997	-0.128	-0.840	0.584
	-	Inoc_Boost	1.235	0.771	120.000	1.602	.499	0.694	-0.168	1.556
D2	-	Inoc	0.164	0.522	120.000	0.314	.998	0.092	-0.488	0.672
	-	Inoc_Boost	1.627	0.676	120.000	2.406	.121	0.914	0.153	1.675
Inoc	-	Inoc_Boost	1.463	0.779	120.000	1.877	.335	0.822	-0.051	1.695

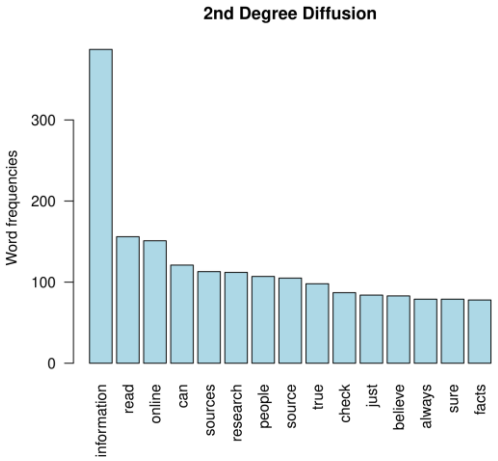
Only Effect Left at T21–T30 is for Inoculation Booster Condition, No Significant Effect for Diffusion Message Exposure

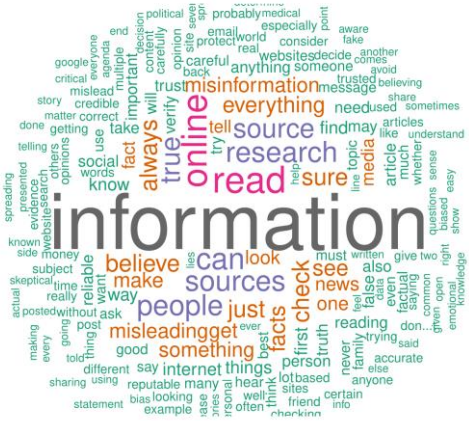


Manipulativeness Discernment

2nd Degree Diffusion (D2)
d = 0.219, p_{tukey} = .857
95% CI [-0.218, 0.657]

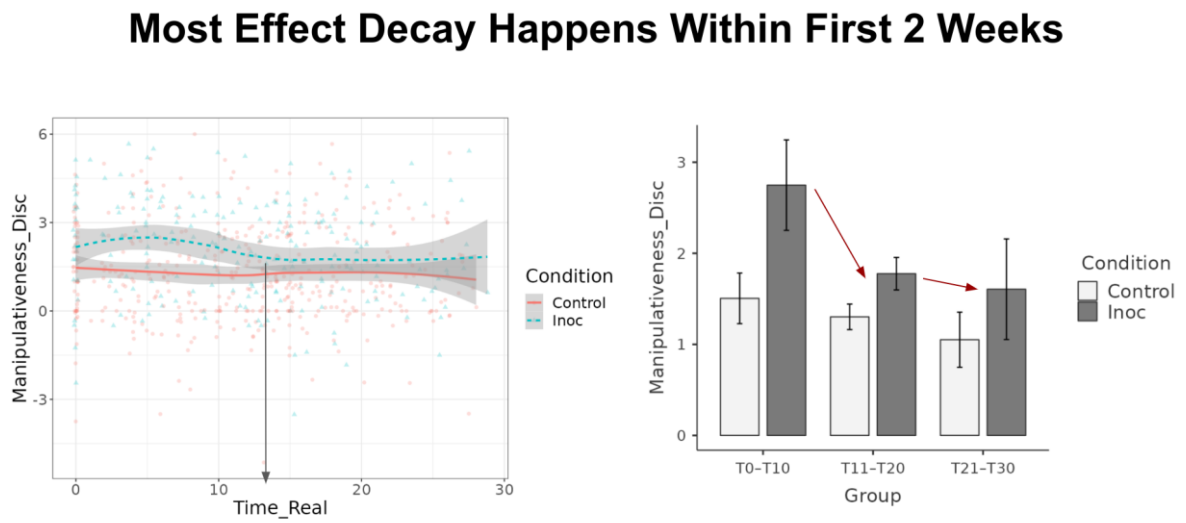
Inoculation Booster (Inoc_Boost)
***d* = 1.133*, p_{tukey} = .037**
95% CI [0.342, 1.925]





Inoculation Effect Decay Curve

To test our decay hypothesis, we expect stronger inoculation effect decay between T0–T10 and T11–T20, than between T11–T20 and T21–T30. We indeed find that the decay is stronger between the first two dates (*d*: 0.756 - 0.286 = -0.470, from significant effect tot non-significant effect), compared to the latter (*d*: 0.286 to 0.334 = 0.048, no significant change, both effects not significant; **evidence for H3**).



Diffusion Message Comparison

Despite the limited evidence for the effectiveness of diffusion messages in this experiment, there are various studies that show that post-inoculation talk may exist and help (Dillingham & Ivanov, 2016; Ivanov et al., 2012). The lack of diffusion effects in this study could be in part due to the artificial setting, and in part due to the limited sample size for small effects. Therefore it is still relevant to look into the content of the diffusion messages to figure out whether participants are able to convey any of the knowledge of the content of the intervention to others, and whether this content is able to spread to a second degree of diffusion or not.

In the table and figure below – also visualised as word clouds and frequency bar graphs in the two sections above – you can see the most frequent words within the diffusion messages. In bold you can see which words are relevant for tackling misinformation, highlighted in blue you can find the words that are directly related to the inoculation intervention topic “vaccines” and the four chapters: (1) emotional storytelling, (2) pseudoscience and fake expertise, (3) naturalistic fallacies, and (4) conspiracy theories. Both diffusion messages show that participants are able to create messages that have something relevant in them with regards to media literacy: many of them are referring to “check sources” and “do a fact check”. However, in the 1st degree diffusion messages, people are also mentioning the topics of the intervention, with messages referring to storytelling, appeal to emotion, and conspiracy theories. These indicators are however no longer present in the 2nd degree diffusion messages. This indicates that the spread of the intervention content of *Bad Vaxx* is unlikely to spread beyond a second person.

1st Degree Diffusion	Frequency	2nd Degree Diffusion	Frequency
information	971	information	387
people	787	**read**	156
facts	626	online	151
can	484	can	121
misleading	441	**sources**	113
research	428	**research**	112
online	376	**people**	107
make	344	**source**	105
look	296	**true**	98
just	294	**check**	87
use	293	just	84
sources	286	believe	83
misinformation	280	always	79
believe	262	sure	79
always	261	**facts**	78
will	235	see	76
read	233	everything	76
check	231	make	69
get	219	**misleading**	65
see	202	something	63
also	200	get	63
vaccines	195	**misinformation**	61
emotional	194	one	57
conspiracy	192	**look**	55
true	180	news	52
media	175	tell	49
sure	170	fact	49
someone	170	**media**	49
important	162	know	48
tell	161	internet	47
things	161	first	46
false	159	things	42
everything	157	find	42
know	157	way	42
try	155	**verify**	41
like	153	truth	40
stories	152	will	39
source	149	try	39
social	145	also	38
fact	145	anything	38
something	143	trust	38
person	142	false	38
way	139	**reading**	37
theories	138	take	36
others	137	**social**	36
say	135	person	36
emotions	135	may	36
one	135	**reliable**	35
many	133	need	35
post	133	article	34

Discussion

In this work we investigated the long-term effectiveness of a 15-minute gamified inoculation intervention that trains people to identify four techniques often used in online vaccine misinformation, including (1) emotional storytelling, (2) pseudoscience and fake expertise, (3) naturalistic fallacies, and (4) conspiracy theories. Specifically, we looked at the long-term effectiveness of such interventions, and whether or not participants are able to craft short inoculation messages to protect their peers against misinformation and therefore spread the inoculation effect beyond those inoculated. We found that the inoculation effect in general stays intact for up to about 10 days, but then is no longer significant unless a booster inoculation (repeating the intervention) is administered. Participants exposed to a first degree diffusion message show descriptively better performance than the control group, but if there is an effect, it was too

small to capture with the current dataset. Comparing first degree and second degree diffusion messages shows that participants are able to craft inoculation messages that talk about the content of the intervention, but that their peers in turn do not create further inoculation messages that include this content. It is therefore unlikely that practitioners can rely on diffusion messages alone to spread protection against misinformation.

Conclusion

The research shows that the Bad Vaxx inoculation effects are replicable and strong, and that it therefore constitutes an excellent addition to any toolkit of counter-misinformation interventions. However, it also shows that alone, if not boosted, the effects decay to insignificance within weeks. This shows the importance of inoculation "booster sessions" to further increase the longevity of the intervention effects. While participants were able to write post-inoculation diffusion messages, they did not suffice to show significant indirect inoculation effects. This therefore suggests that it is not sufficient to rely on post-inoculation talk alone as a method to diffuse resistance to persuasion.

Next steps

While this research does not provide support for the effectiveness of post-inoculation-talk-based diffusion, previous research has shown that post-inoculation talk can have benefits (Dillingham & Ivanov, 2016; Ivanov et al., 2012). These contradictions should be further explored, for example in designs with larger sample sizes and concentrated over shorter time frames to establish a baseline effect. In addition, other post-inoculation evaluation formats could be tested instead of the current type of diffusion messages. For example, participants could be asked to respond to specific arguments or misleading stimuli in order to see how they explain what is misleading. This explanation could then be shown to others, rather than "general" diffusion messages.

Supplementary materials

All supplementary materials for this project, including the cleaning script, dataset, diffusion messages, and Qualtrics surveys, can be found on the OSF for this study at https://osf.io/9a5en/?view_only=8f2b8b7a04ad48bd99fa88c922d08d42 (in the "Study 2" folder).

References

Capewell, G., Maertens, R., Linden, S. van der, & Roozenbeek, J. (2023). *Misinformation interventions decay rapidly without an immediate post-test*. PsyArXiv. https://doi.org/10.31234/osf.io/93ujx

Dillingham, L. L., & Ivanov, B. (2016). Using postinoculation talk to strengthen generated resistance. *Communication Research Reports*, *33*(4), 295–302. https://doi.org/10.1080/08824096.2016.1224161

Ivanov, B., Miller, C. H., Compton, J., Averbeck, J. M., Harrison, K. J., Sims, J. D., Parker, K. A., & Parker, J. L. (2012). Effects of postinoculation talk on resistance to influence. *The Journal of Communication*, *62*(4), 701–718. https://doi.org/10.1111/j.1460-2466.2012.01658.x

Leder, J., Schellinger, L., Maertens, R., Chryst, B., Linden, S. van der, & Roozenbeek, J. (2023). *Feedback exercises boost discernment and longevity for gamified misinformation interventions*. PsyArXiv. https://doi.org/10.31234/osf.io/7k2mt

Lewandowsky, S., & van der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, *32*(2), 348–384. https://doi.org/10.1080/10463283.2021.1876983

Lu, C., Hu, B., Li, Q., Bi, C., & Ju, X.-D. (2023). Psychological inoculation for credibility assessment, sharing intention, and discernment of misinformation: Systematic review and meta-analysis. *Journal of Medical Internet Research*, *25*, Article e49255. https://doi.org/10.2196/49255

Maertens, R., Roozenbeek, J., Simons, J., Lewandowsky, S., Maturo, V., Goldberg, B., Xu, R., and van der Linden, S. (2023). *Psychological Booster Shots Targeting Memory Increase Long-Term Resistance Against Misinformation*. PsyArXiv. https://doi.org/10.31234/osf.io/6r9as