
Toolbox of individual-level interventions against online misinformation

In the format provided by the
authors and unedited

Supplementary Information for the Toolbox of Individual-Level Interventions Against Online Misinformation

A. Selection of Interventions

Criteria for Inclusion of Interventions:

To be included in the toolbox, intervention types must meet five criteria.

1. Definition and scope: The intervention type must be a well-defined behavioral or cognitive approach that targets individual behavior, beliefs, attitudes, and competences. Interventions at a systemic level (e.g., content moderation, algorithmic filtering) and full educational curricula are not within the scope of this toolbox.

2. Problem addressed: The intervention type must explicitly address the challenge of misinformation, encompassing issues such as disinformation, false and misleading information, fake news, and related issues.

3. Nonredundancy: The intervention type should not significantly overlap with others in the toolbox. It should offer a unique and distinct contribution to combating misinformation.

4. Evidence: There must be empirical studies with published findings available in English that support the efficacy of the intervention. The strength of the evidence should be sufficient; this may include replication studies or multiple publications supporting the intervention, depending on the context.

5. Expert opinion: All papers recommended for inclusion in the evidence toolbox were suggested by the experts who co-authored this toolbox, drawing upon their extensive knowledge and experience in the field.

Based on the first four criteria, coordinating authors re-evaluated existing interventions in the toolbox and decided to remove “Self-reflection tools” on the basis that it does not satisfy the second inclusion criterion (it addresses microtargeted manipulation instead of misinformation) and to merge “Debunking” and “Rebuttals of science denialism” under “Debunking and rebuttals” on the basis of the criterion of nonredundancy. The selected nine types of interventions include: accuracy prompts, debunking and rebuttals, friction, inoculation, lateral reading and verification strategies, media-literacy tips, social norms, source-credibility labels, and warning and fact-checking labels.

The experts on these interventions (i.e., researchers who have published work on the topic; see part C of this SI) then identified relevant information on the conceptual details and empirical evidence underlying them. To ensure that our evidence base spans a wide variety of samples and populations, we also actively approached experts in misinformation interventions from underrepresented regions. They nominated an additional set of research papers for all interventions types.

To ensure the completeness of intervention types in the toolbox, we analyzed additional intervention types from relevant projects: the Mercury Project Misinformation Intervention Comparison Study [1] and the systematic review conducted by [2]. Within this process, we considered the following intervention types.

Source-credibility labels (labels added to information source): While most studies on source credibility focus on the source issuing corrections [3, 4] or on highlighting the publisher [5], there has been limited exploration of adding credibility labels to content. Recently, three recent studies delved into this: Aslett et al. [6] tested whether flagging the source as trustworthy or untrustworthy (using NewsGuard ratings) would impact false news consumption. They found no significant effects of the intervention on the general population but noted an improvement in the information consumption habits of the heaviest misinformation consumers. Celadin et al. [7] tested a star-rating intervention for news sources and found that this approach enhanced participants’ discernment towards sharing, specifically by reducing the intention to share false headlines. Finally, Prike et al. [8] implemented “credibility badges” in a social-media simulation. These badges associated accounts, including participants’, with a credibility score. Participants’ credibility score changed dynamically based on how they engaged with true and false posts. The badges were found to improve discernment of true and false claims and improve engagement with misleading posts (i.e., more flagging and less liking and sharing). Although the overall evidence on this intervention type is mixed, it is evolving, and source-credibility labels are already in widespread use on social media. We therefore included source-credibility labels in the toolbox.

Crowdsourced ratings: While there are studies that focus on the efficacy [9] and potential partisanship of crowdsourced accuracy ratings [10], to our knowledge, there are no empirical papers assessing the individual-level impact of this intervention type. A recent preprint [11] examines the “misleading count” intervention, which does not fit neatly into existing intervention types. In practice, X (formerly Twitter) has introduced Community Notes (<https://help.twitter.com/en/using-x/community-notes>), a cooperative community-based intervention that enables users to contextualize misleading information. However, this intervention does not meet the criteria for inclusion in the toolbox due to the absence of experimental evidence.

Lateral reading nudge: Among the three papers reviewed by Ziemer and Rothmund [2], the paper by Donovan and Rapp [12] is already included in our toolbox under lateral reading and verification strategies. Thornhill et al.’s [13] “BalancedView” intervention, while interesting, does not fully align with the definition of lateral reading. Kobayashi et al.’s [14] intervention, an instruction to search the Web, does not genuinely constitute a nudge. We have included it under lateral reading and verification strategies. The lateral reading nudge can be subsumed under the same intervention type and therefore does not meet the criteria for inclusion in the toolbox.

Self-affirmation (referred to as identity management in [2]): This intervention type primarily complements corrections and is typically administered alongside corrections to promote greater responsiveness to corrective information. Consequently, it does not meet the criteria for inclusion in the toolbox.

Social corrections (referred to as social invalidation in [2]): Social correction is an integral feature of debunking and does not represent a distinct intervention type.

Various media and science literacy studies: In accordance with our first criterion, we do not include entire media literacy curricula or courses in our study. However, we are open to considering specific components of educational interventions that can be regarded as boosts. For instance, lateral reading is included in our toolbox (but not the

complete Civic Online Reasoning curriculum). Ziemer and Rothmund’s review [2] also encompasses a range of studies exploring the effects of media and science literacy. Most of these are unique, examining the impact of specific structured courses, infographics, or guidelines, but they do not represent a specific intervention type. Media-literacy games might be considered an exception, but again, they seem to be more about the format of implementation rather than a distinct intervention type. To include relevant studies, we expanded “lateral reading” to “lateral reading and verification strategies” and included several new papers, such as one on the Misinformation Is Contagious game, which offers strategies for assessing online information and responsible sharing [15] and one on the online educational game Trustme!, designed to foster evaluation strategies [16].

B. Online Supplement

The toolbox has an online supplement in the form of two dynamic tables: a [conceptual overview](#) and an [evidence overview](#). The online supplement also contains [selected examples of interventions](#) and [a world map of evidence](#). The dynamic tables were built with the R package *reactable* [17]. The website was built using the *workflowr* package [18].

C. Co-authors and Their Areas of Expertise

Co-authorship of the toolbox follows a two-tier structure (similar to the Debunking Handbook; see [19]), with coordinating and contributing authors. Contributing authors are mainly experts on different types of interventions. They were invited based primarily on one key criterion: Having published scientific work testing the effectiveness of cognitive and behavioral interventions against misinformation. A few co-authors who do not work on specific interventions contributed to other aspects of the manuscript and the toolbox (e.g., description of the field studies paradigm or discussion on combining interventions). In addition, in the second stage of developing the toolbox, we invited four experts on misinformation interventions from Pakistan, Mexico, China, and Israel in order to address the limited representation of samples beyond North America and Europe. We do not claim that our expert team is representative of the entire field, but we believe it offers good coverage nevertheless. For each co-author’s contribution, see Table S1.

Table S1: Co-authors

Name	Affiliation	Country	Contribution to the toolbox	Academic discipline
Anastasia Kozyreva	Max Planck Institute (MPI) for Human Development	Germany	General coordination; cognitive and behavioral interventions	Philosophy and cognitive science
Philipp Lorenz-Spreen	MPI for Human Development	Germany	General coordination; cognitive and behavioral interventions	Network science
Stefan Herzog	MPI for Human Development	Germany	General coordination; cognitive and behavioral interventions	Psychology
Ullrich Ecker	University of Western Australia	Australia	General coordination; debunking and rebuttals, source-credibility labels, social norms	Psychology
Stephan Lewandowsky	University of Bristol; University of Western Australia, University of Potsdam	UK, Australia, Germany	General coordination; inoculation, debunking and rebuttals	Psychology
Ralph Hertwig	MPI for Human Development	Germany	General coordination; cognitive and behavioral interventions	Psychology
Ayesha Ali	Lahore University of Management Sciences	Pakistan	Media-literacy tips, global studies of misinformation interventions	Economics
Joe Bak-Coleman	Columbia University	USA	Discussion on combining interventions and collective behavior	Collective behavior
Sarit Barzilai	University of Haifa	Israel	Lateral reading and verification strategies; global studies of misinformation interventions	Educational psychology
Melisa Basol	University of Cambridge	UK	Inoculation	Psychology
Adam Berinsky	Massachusetts Institute of Technology	USA	Warning and fact-checking labels	Political science

Name	Affiliation	Country	Contribution to the toolbox	Academic discipline
Cornelia Betsch	University of Erfurt	Germany	Debunking and rebuttals	Psychology
John Cook	University of Melbourne; George Mason University	Australia; USA	Debunking and rebuttals, inoculation	Psychology
Lisa Fazio	Vanderbilt University	USA	Friction	Psychology
Michael Geers	MPI for Human Development	Germany	Description of the paradigm ‘field studies on social media’	Psychology
Andrew Guess	Princeton University	USA	Media-literacy tips	Political science
Haifeng Huang	Ohio State University	USA	Debunking and rebuttals, global studies of misinformation interventions	Political science
Horacio Larreguy	Instituto Tecnológico Autónomo de México	Mexico	Inoculation, global studies of misinformation interventions	Economics
Rakoen Maertens	University of Oxford	UK	Inoculation	Psychology
Folco Panizza	IMT School for Advanced Studies Lucca	Italy	Lateral reading and verification strategies	Psychology
Gordon Pennycook	Cornell University; University of Regina	USA; Canada	Accuracy prompts	Psychology
David G. Rand	Massachusetts Institute of Technology	USA	Accuracy prompts	Psychology
Steve Rathje	New York University	USA	Inoculation	Psychology
Jason Reifler	University of Exeter	UK	Warning and fact-checking labels	Political science
Philipp Schmid	University of Erfurt	Germany	Debunking and rebuttals	Psychology

Name	Affiliation	Country	Contribution to the toolbox	Academic discipline
Mark Smith	Stanford University	USA	Lateral reading and verification strategies	Educational psychology
Briony Swire-Thompson	Northeastern University	USA	Debunking	Psychology
Paula Szewach	University of Exeter; Barcelona Supercomputing Center	UK; Spain	Warning and fact-checking labels	Political science
Sander van der Linden	University of Cambridge	UK	Inoculation	Psychology
Sam Wineburg	Stanford University	USA	Lateral reading and verification strategies	Educational psychology

References

- [1] Susmann, M., Fazio, L., Rand, D.G., Lewandowsky, S.: Mercury Project Misinformation Intervention Comparison Study (2023) <https://doi.org/10.17605/OSF.IO/FE8C4> . Preprint.
- [2] Ziemer, C.-T., Rothmund, T.: Psychological underpinnings of misinformation countermeasures. *J. Media Psychol.* (2024) <https://doi.org/10.1027/1864-1105/a000407>
- [3] Ecker, U.K.H., Antonio, L.M.: Can you believe it? an investigation into the impact of retraction source credibility on the continued influence effect. *Mem. Cognit.* **49**(4), 631–644 (2021) <https://doi.org/10.3758/s13421-020-01129-y>
- [4] Vraga, E.K., Bode, L.: Using expert sources to correct health misinformation in social media. *Sci. Commun.* **39**(5), 621–645 (2017) <https://doi.org/10.1177/1075547017731776>
- [5] Dias, N., Pennycook, G., Rand, D.G.: Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harv. Kennedy Sch. Misinformation Rev.* (2020) <https://doi.org/10.37016/mr-2020-001>
- [6] Aslett, K., Guess, A.M., Bonneau, R., Nagler, J., Tucker, J.A.: News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Sci. Adv.* **8**(18) (2022) <https://doi.org/10.1126/sciadv.abl3844>
- [7] Celadin, T., Capraro, V., Pennycook, G., Rand, D.G.: Displaying news source trustworthiness ratings reduces sharing intentions for false news posts. *J. Online Trust Saf.* **1**(5) (2023) <https://doi.org/10.54501/jots.v1i5.100>
- [8] Prike, T., Butler, L., Ecker, U.K.H.: Source-credibility information and social norms improve truth discernment and reduce engagement with misinformation online (2023) <https://doi.org/10.31234/osf.io/dhx6f> . Preprint.
- [9] Pennycook, G., Rand, D.G.: Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. U. S. A.* **116**(7), 2521–2526 (2019) <https://doi.org/10.1073/pnas.1806781116>
- [10] Allen, J., Martel, C., Rand, D.G.: Birds of a feather don’t fact-check each other: Partisanship and the evaluation of news in twitter’s birdwatch crowdsourced fact-checking program. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–19 (2022). <https://doi.org/10.1145/3491102.3502040>
- [11] Pretus, C., Javeed, A.M., Hughes, D.R., Hackenburg, K., Tsakiris, M., Vilarroya, O., Bavel, J.J.V.: The Misleading count: An identity-based intervention to counter partisan misinformation sharing. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **379**,

20230040 (2024) <https://doi.org/10.31234/osf.io/7j26y>

- [12] Donovan, A.M., Rapp, D.N.: Look it up: Online search reduces the problematic effects of exposures to inaccuracies. *Mem. Cognit.* **48**(7), 1128–1145 (2020) <https://doi.org/10.3758/s13421-020-01047-z>
- [13] Thornhill, C., Meeus, Q., Peperkamp, J., Berendt, B.: A digital nudge to counter confirmation bias. *Front. Big Data* **2** (2019) <https://doi.org/10.3389/fdata.2019.00011>
- [14] Kobayashi, T., Taka, F., Suzuki, T.: Can “googling” correct misbelief? Cognitive and affective consequences of online search. *PLoS One* **16**(9), 0256575 (2021) <https://doi.org/10.1371/journal.pone.0256575>
- [15] Barzilai, S., Mor-Hagani, S., Abed, F., Tal-Savir, D., Goldik, N., Talmon, I., Davidow, O.: Misinformation is contagious: Middle school students learn how to evaluate and share information responsibly through a digital game. *Comput. Educ.* **202**, 104832 (2023) <https://doi.org/10.1016/j.compedu.2023.104832>
- [16] Yang, S., Lee, J.W., Kim, H.-J., Kang, M., Chong, E., Kim, E.-m.: Can an online educational game contribute to developing information literate citizens? *Comput. Educ.* **161**, 104057 (2021) <https://doi.org/10.1016/j.compedu.2020.104057>
- [17] Lin, G.: Reactable: Interactive Data Tables for R. (2024). R package version 0.4.4.9000, <https://github.com/glin/reactable>. <https://glin.github.io/reactable/>
- [18] Blischak, J.D., Carbonetto, P., Stephens, M.: Creating and sharing reproducible research code the workflowr way. *F1000Res.* **8**, 1749 (2019) <https://doi.org/10.12688/f1000research.20843.1>
- [19] Lewandowsky, S., Cook, J., Ecker, U.K.H.: Under the Hood of *The Debunking Handbook 2020*: A consensus-based handbook of recommendations for correcting or preventing misinformation (2020). <https://skepticalscience.com/docs/DebunkingHandbook2020-UnderTheHood.pdf>