PRSS Pakistan Review of Social Sciences (PRSS)
ISSN : 2708-0951

ORGINAL ARTICLE

How correction formats change belief updating in misinformation exposure across digital platforms

Muhammad Faraz Arif

Fake News Watchdog (FNW)

Abstract

Digital platforms have increased sources of misinformation exposure and tools to counteract the misinformation available to us, but the updating of beliefs is skewed. The article focuses on the role of correction format in influencing the belief revision following exposure to misinformation in platform conditions. The study utilized a secondary qualitative research design to synthesize qualitative results of published studies on digital misinformation correction to derive an inductive, mechanism-based explanation of why and when individuals update beliefs, do not update beliefs, or change behavior, instead of beliefs, to arrive at its result. There are five themes that the synthesis brings out. To begin with, when credit is expressed with the help of format cues (transparency, sourcing, and perceived legitimacy), format cues are also able to be counter-argued when they are used as a partisan contention. Second, there is an enormous conditioning effect of cognitive effort; in-feed labels and micro-prompts have a high frequency of encounter but a shallow update. In contrast, narrative corrections and detailed explainers have a high capacity to produce deeper updating but low uptake in attention-sparse conditions. Third, identity threat and reactance arise when the corrections are seen as coercive, especially when the politicized deals are concerned or when there are signals that relate to moderation, investigating the censorship frames. Fourth, conversational conditions repurpose the result: thread corrections, when used to encourage learning among bystanders via social proof, can actually promote learning, but may also intensify polarization and foster misinformation via repetition. Fifth, timing and visibility moderate effectiveness: prebunking may influence the interpretation in advance when familiarity has not yet solidified, and debunking works best when it has coherent alternative explanations and when it reaches the users early enough to counter repeated exposure. In general, the results indicate that the conditional factors of correction format according to the affordances of the platform, social meaning, and interventions should be assessed on both examples of belief and behavioral outcomes. The article rounds off with design suggestions regarding platforms and research recommendations on the persistence, individually conveyed messaging conditions, and developing correction facilities.

Keywords: misinformation; corrections; belief updating; qualitative synthesis; fact-checking; prebunking; warning labels; community notes; platform affordances; reactance; cognitive load; information credibility

Introduction

The effect of misinformation on online sources tends to remain after correction due to the fact that changing beliefs is not as simple as substituting false beliefs with real-life ones. Studies into the continued influence effect reveal that retractions often fail when they cause the removal of a claim without giving an alternative reason that can enable people to make sense of the incident (Lewandowsky et al., 2012). This issue is aggravated in social and video media because people are exposed to misinformation in certain environments. Feeds are facilitating fast, low-effort consumption, whereas search interfaces are promoting more thoughtful consumption, whilst the tools used to conduct private messaging can help seal claims off to open debate. These differences are important in that they influence the recognition of corrections, their in-depth processing, and retention in memory among the users. That is why it is convenient to consider the unit of analysis to be a format. The above corrections vary by time of attack (pre-exposure-based protection with before exposure and after-attack-based protection with after-exposure), in cues (experts, peers, or system-made notifications), in medium (labels, fact-check cards, videos, or comment responses), in force (weak warnings versus removals), and setting (linked directly to the misinformation versus stored away as distinct links). Each of these design decisions has the potential to alter perceptions of credibility and cognitive effort needed, which, in turn, can alter the presence or absence of belief updating.

The multidimensional definition of belief updating should also be present. It may include the changes in the perceived accuracy and confidence, the changes in the willingness to share, the decrease in the further impact of the false statement, and the alterations of the trust in the sources or platforms (Lewandowsky et al., 2012). Such results do not necessarily move jointly. Other techniques decrease sharing by increasing friction or reputational warnings, although individuals may still have retained partially the same private beliefs. These processes are also influenced by social context when the misinformation is mixed up with identity and allegiance to groups. Preliminary studies in the field of political psychology have established that some corrections fail in highly committed groups and occasionally some misperceptions are in part enhanced in the case of experimental corrections that are experienced as a threat (Nyhan and Reifler, 2010). Subsequent findings indicate that the severe effects of backfire are less prevalent than initially believed. However, opposition to correction is still manifested under certain conditions, especially in situations involving activation of identity-protective reasoning (Wood and Porter, 2019). It means that format is not only important as to the information that is conveyed, but can diminish or increase defensiveness.

Several studies demonstrate how formats specific to platforms can influence processes of belief updating. Corrective information that is conveyed with the help of interface features, introducing it close to a misleading claim, may decrease misperceptions, reducing costs of search and focusing attention on more quality-oriented sources (Bode and Vraga, 2015). Even the delivery of corrections in social situations may impact observers as opposed to the original poster. As one example, professional corrections in comment boards have been demonstrated to diminish faith in health misinformation among comment board readers, in part due to the high credibility signal produced by expertise and the fact that so much elucidation does not need to be provided (Vraga and Bode, 2017). Another format is prebunking, which prepares users to perceive manipulation methods to minimize future responses to false information through the interpretation of new statements (van der Linden et al., 2017). Meanwhile, meta-analyses of the existing state of literature of fact-checking reveal that corrections have positive average effects, which, however, are different across audience features, topics, and message structures, and postulate the necessity to clarify what and whom works, and why (Walter et al., 2020). This motivates the present study's focus on the following question:

How do correction formats shape mechanisms of belief updating after exposure? How do platform affordances and social settings moderate these effects, and which recurring barriers, such as distrust and identity threat, lead to partial updating or persistent misperceptions?

Conceptual Background

The post-exposure updating of beliefs following misinformation is influenced by the interaction of the cognitive, social, and platform forces and not information deficit alone. Repeated exposure boosts processing fluency and familiarity at the cognitive level, which is prone to confusion with truth, even by knowledgeable users, to provide a structural advantage to repeatedly seen misinformation in feeds of high velocity (Fazio et al., 2015). This complicates the process of correction since individuals can subsequently recover the "gist" of the claim but forget the refutation, which continues influencing the effect unless the correcting information provides a consistent alternative explaining the stance that can replace the initial mental model (Lewandowsky et al., 2012; Prike and Ecker, 2023). Online space magnifies such insecurities by adding repetition, attention dispersal, and time to reflect. Consequently, formats requiring some extra effort (long explainers, external links) might be conceptually perfect and used sparingly. In contrast, low-effort formats (labels, banners) might be read but shallowly processed, generating partial instead of lasting change. Another important implication that can be made is that format is not merely a delivery option, but rather a mechanism that alters cognitive load, retrieval conditions, and the availability of a usable alternative narrative per se.

Identity-relevant reasoning is another limitation of motivated reasoning that limits the effects of correction. Enigmatically, individuals can tendentially give congenial credit and disregard threatening modifications to information in polarized situations, and greater reasoning power can occasionally encourage selective processing since analytical resources are brought to bear in support of identity-supportive findings (Kahan et al., 2017). This is not to say corrections do not work; it is to state that the road to updating is conditional in regard to the perceived threat of the correction and whoever happens to be administering it. The early political literature emphasized situations in which corrections were ineffective and sometimes tended to create more misperceptions between dedicated subgroups (Nyhan and Reifler, 2010). In contrast, later extensive evidence indicates dramatic patterns of backfire are rare, but that resistance and little updating can still be important boundary conditions (Wood and Porter, 2019). The most policy-relevant fact is that format can moderate defensiveness: punitive or stigmatizing cues might invite reactance, and formats that respect face, minimise humiliation in the editing section, and offer alternative, respectful explanations may be helpful to update without increasing identity threat.

Spread and correction uptake are both products of affective mechanisms. Emotionally stimulating content, particularly moral outrage, spreads quickly in social networks and may lead to the development of an attention asymmetry when false information spreads more and faster than corrections (Brady et al., 2017). Even accurate corrections that are effectively inert may be at a disadvantage in the case of contests. This is important to make a format decision on video-first platforms where the style of persuasion, narrative speed, and trustworthiness of creators can make the difference between what gets viewed and what gets retained. There is also a strong emphasis on trust and credibility heuristics: in time-constrained situations, users have a tendency to use such cues as perceived expertise, independence, transparency, and consistency instead of analyzing evidence appropriately (Metziger and Flanagin, 2013). Certain formats, like fact-check cards with sources, community notes with visible reasoning, foreground their transparency, whereas others, like generic labels, removals without explanation, can make some communities distrustful. Thus, factual correction of the same type that results in various outputs can be achieved through the format conveying procedural fairness and epistemic humility in lieu of the format indicating institutional power.

These families of mechanisms provide incentive to a typology of formats of correction in the form of interventions that manipulate attention, credibility cues, and processing costs. In the case of warning labels and accuracy prompts, they are also fast heuristics that can minimize endorsement and sharing, where effects are measured by visibility and source credibility (Martel & Rand, 2023). Fact-check cards and other related article modules seek to reduce the cost of verification by making the alternative information more available at the moment of exposure; they have constraints of uptake and bias avoidance. Prebunking (inoculation) is structurally different in that it aims at manipulating recognition and interpretive resilience as opposed to refuting individual claims, which can be beneficial in high-volume settings with which debunking cannot scale claim-by-claim (Lewandowsky and van der Linden, 2021). Community-based corrections place authority within cross-user consensus and circumstances, and also may bring about credibility challenges, particularly in polarization, and their effect relies crucially on whether the added context becomes observable before resharing accelerates (Renault et al., 2024). The sharing discernment of friction-based formats (read-before-share, accuracy prompts) can enhance the ability to change attention to accuracy at the point of decision, even when beliefs remain unchanged, which is noteworthy since not everyone sharing misinformation reports does so due to a conviction but by default (Pennycook et al., 2022). Lastly, removals and strikes are powerful signals and alerts. However, they can intensify the censorship of stories in cases of low transparency, so they represent a high-stakes format choice in contentious responsibility areas.

Methodology

A secondary qualitative research design with an inductive approach of analysis was used in this study due to the intention not to test a predetermined causal model but to explain how correction formats influence belief updating in diverse digital contexts. The inductive method was suitable since platform correction practices (e.g., labels, friction prompts, community annotations, removals) are heterogeneous and dynamically changing, and so strict a priori coding frames may run the risk of introducing old presumptions on modern phenomena (Braun & Clarke, 2006). Rather, it was intended to allow explanatory patterns to emerge out of the qualitative evidence, and to be sensitive to how desirable results on platforms are mediated by platform affordances and social interaction to account for the meaning of correction in action. The secondary qualitative research approach was especially appropriate since, in this way, it would be possible to synthesize data between several settings and populations, with the study consequently recognizing repeat mechanisms and organizing circumstances that should not be triggered in single-site qualitative research efforts. In an area where misinformation and correction tools differ by platform and subject, it would be advantageous to qualify findings of the studies through synthetic approaches; nevertheless, retaining sensitivity to context and meaning-making (Thomas & Harden, 2008).

The process of finding studies was carried out using systematic searches of large interdisciplinary databases of communication and platform research used, such as Scopus, Web of Science, Communication & Mass Media Complete, and ACM Digital Library. The sampling was narrowed down to recent years to indicate the current platform governance and correction infrastructures. The identified studies included those that investigated misinformation correction in an online setting and provided qualitative results (either an independent, qualitative study or mixed-method research with results in extractable qualitative form) on the format of correction and the sense-making that users apply to belief updating. omitted studies did not have substantial substance on formats of correction or offered no qualitative information on the interpretations of users, acceptance, resistance, or sharing in decision-making. The information was programmed into a structured template that captured contexts of the platform, correction format features (timing, modality, source cues, embedding, and strength), and the reported outcomes involved in terms of belief updating (e.g., change in the perceived accuracy or confidence, share willingness, trust judgments, and persistence of false claims).

Analytical induction occurred within inductive coding of the derived results, followed by grouping of codes into descriptive themes and subsequently formulating higher-order analytic themes which explained how and why certain forms were successful in updating or creating resistance in certain platform situations (Braun and Clarke, 2006). Synthesis The synthesis focused on constant comparison across platforms and formats to prevent overgeneralization and find negative cases that could not be corrected or reacted to. The credibility was enhanced through keeping an audit trail of the coding decisions, writing reflexive memories about assumptions on the truth, governing authority, and the platform governing, and evaluating faith in themes by reflecting on coherence between studies, sufficient supporting evidence, and appropriateness to the focus of the review (Lewin et al., 2015).

Findings

Credibility cues

The credibility was not, in studies, a fixed attribute of truth (that is, of the truth) but an inference that users drew out of format indicators, including sourcing, transparency, and identity of whoever purported to have corrected. Evidence, connected sources, and described reasoning in corrections tended to make them more accepted due to the reduction of uncertainty and the indication of procedural fairness (Ecker et al., 2022). Meanwhile, detailed formats also provided more surface on which to motivate rebuttal. The respondents were free to question those elements of the explanation selectively, cast doubts on the origin, or interpret the correction as partisan talking as opposed to clarification. This trend was particularly evident with claim-by-claim fact-check cards and long explainers that were sometimes irresistibly dense with information and open to counter-dismissal. The community-based corrections gave the institutions less power and placed more weight on negotiated legitimacy. In cases where users perceived community context in a cross-partisan and transparent way, credibility increased. In contrast, in cases where they interpreted it as either coordinated or biased, the leading paradigm was that of legitimacy challenges. Community Notes: The presence of large platform-level evidence and observation points to the idea that attached context can soften the dissemination of erroneous content, yet it only relies on exposure. Influences were greatest when notes were made visible early enough to influence sharing choices (Slaughter et al., 2025). Platform-scale studies are stronger in terms of their ecological validity but susceptible to selection and timing confounds, as posts that receive notes are not the same as those that do not receive notes. The common mode of belief updating in this case was visible attribution, less uncertainty, acceptance, or less resharing. The fallacy used was: undermining of credibility, undermining of the source, and perpetuated use of the initial assertion.

Effort and friction

One recurring conclusion was that many corrections are not passed due to unprocessed corrections, and not due to incorrect ones. In-feed labels and short warnings, which are a low-effort format, maximize reach, although they usually result in apathetic updating. Users can reduce their confidence to a low level or even not share, but the mental representation may be preserved. More intense forms like click-through explainers and more lengthy narrative de-bunks have higher potential to be comprehended and replace memories, yet have low uptake in attention-constrained settings (Okuhara et al., 2025). This is a fundamental platform design trade-off: there are trade-offs between interventions that are depth-optimized and interventions that are exposure-optimized; and there are trade-offs between interventions that are cognitive-integration-optimized and interventions that are distribution-optimized. Evidence of accuracy makes clear the reason why some friction tools continue to be effective even when the beliefs remain hardly in motion. According to drift diffusion modelling, the use of prompts can shift weighting by individuals at the point of posting instead of enhancing deliberation in general and making judgments more accurate without enduring belief correction (Lin et al., 2023). Methodologically, much of the work on friction has been based on short follow-up windows, which do not do much to justify claims that people update beliefs over time, but are helpful in decoupling behavior change and belief change. The average trajectory was as follows: small interruption, more concern for accuracy, and less impulsive sharing. Habituation to failure, fatigue, or users clicking beyond cues when familiarity with the misinformation is increasing was the failure mode.

Reactance risk

The experiments over and over again demonstrated that correction formats can be construed as a control, which elicits the re-actance and diminishes update, particularly when the conversation is politicized or related to group identity. Hard labels, removals, and strike signals were stronger interventions that were likely to be decoded through censorship frames. In such instances, the users did not consider the accuracy of the claim and judged the platform based on its motives. In contrast, the correction turned into a representation of authority and not information (Ecker et al., 2022). This does not imply that strong formats are necessarily backfiring. It implies that they can only be works depending on their perceived fairness, transparency, and how the users trust an intervening institution. Face-saving corrections and narrative form were frequently said to be less threatening since a user can change opinions without being seen as a loser. Narrative is, however, not a guarantee. A registered report of the comparison

between narrative and non-narrative corrections did not find any consistent narrative advantage in conditions where informativeness was kept constant, indicating that narrative efficacy is construction-dependent, such as the coherence, identification, and efficiency to provide a plausible alternative explanation (Ecker et al., 2020). Preregistered designs enhance credibility methodologically due to the reduction of researcher degrees of freedom. However, controlled experiments may still fail to capture the social incentives that are pervasive on real platforms. This was the usual direction: low identity threat, less counter-arguing, and incorporation of the correction. The failure mode was perceived coercion, reactance, and entrenchment, mostly in the form of rejecting the messenger and not the facts.

Social dynamics

This was the case with corrections that were embedded in conversations, which, unlike standalone ones, generated more than one audience. The target of correction influenced the results, observers in the network, and bystanders. Public thread corrections have the potential to induce social proof and accuracy norms, which benefits bystander updating even in cases where the originator of the information does not yield. Meanwhile, social correction is a source of escalating conflict, amplification of the misinformation by restating it, and performance of the dunking, in particular, on rapid public media where outrage is given attention. A study of WhatsApp and Facebook established that political discourse, interdependence in cross-cutting and relational characteristics, influenced willingness to correct and its outcomes. Correction can be shunned in close connections to minimize conflict, whereas in more mixed networks, correcting can be done, but based on partisan identity (Rossini et al., 2020). Survey studies gather realistic relational constraints of the world, but, methodologically, are frequently reliant on self-report, a factor that tends to obscure actual belief change with self-presentation. The common route consisted of: perceptual/mental correction in a social environment, bystander learning, less sharing, and occasionally corrected judgments. Failure mode was: status competition and polarization, and the other identity signal was correction.

Timing effects

Timing predetermined the success of formats: whether worked out in the initial interpretation or revised. The prebunking interventions were used to inoculate users to manipulation techniques before they were exposed. Inoculation games provide evidence of resistance and confidence, with further refinement of measures and later work, but cross-cultural and cross-topic generalization is still a major weakness (Roozenbeek et al., 2018; Basol et al., 2020). Upstream skills were also found to be relevant since digital media literacy interventions enhanced the ability to differentiate mainstream and fake news in a variety of situations (Guess et al., 2020). On the contrary, post-exposure debunking requires familiarity and memory persistence to be overcome. Research into the effectiveness of correction focuses on the fact that giving an account of the reasons that led to the development of the misinformation, or another version of causality, can enhance correction since it promotes the replacement of mental models coherently as opposed to sharply (Connor Desai & Reimers, 2023). The most important failure mode is familiarity: restating a myth to disprove it may reinforce the recurrence of this myth in the future, in case the disproving tag is not encoded heavily, or the correction tag is forgotten. The common route was: an early frame-setting/strong replacement process explanation, with less dependence on misinformation over time. Failure mode: Repeated exposure, which did not lead to permanent integration, so familiarity favoured the false claim.

Visibility constraints

One last theme was about attention economics and distribution. Communications may be powerful in content but lack in reach. In comparison with the original misinformation, many formats are under-distributed, particularly when aggressive algorithms are at work to maximize engagement as opposed to accuracy. Multi-modal presentation can transform the pathway of persuasion on video channels. Indicatively, experimental studies indicate that audiovisual qualities, including tempo, have the potential to affect counter-arguing and perceived credibility in brief debunking videos (Zhang et al., 2024). Such studies are useful to isolate features of format. Still, they also run the risk of exaggeration when they overlook those things, such as algorithmic replay, trust between creators and fans, and being selected for exposure. The usual channel was: vis-correction, believable signal, and controllable effort, involvement with correction, revision, or diminished sharing. The invisibility failure mode was the absence of engagement, and thus, there was no updating, whether of high quality or not.

Discussion

The results offer implications that correction formats have fewer effects on belief updating by providing more facts and fewer by creating more favorable situations in which individuals judge a claim to be worth updating. The cues of credibility were always centrally placed, and the credibility of a fact was not inseparable from truth. It was concluded based on the cues of the format, like transparency, sourcing, and perceived legitimacy of the actor who is correcting (Ecker et al., 2022). This is the reason why the detailed fact-check cards can both enhance and undermine the results of correction. When users construe the correction as an attempt to act in good faith with regard to epistemics, detail diminishes

uncertainty and facilitates integration. When the users make it into a partisan argument, detail is more controversial and contributes to counter-arguing. Community-based formats partly address the issue of legitimacy, as they decentralize authority. Still, the data suggest that they can also recreate the credibility problem in a new key, particularly in situations where the users doubt the authority to determine what context is acceptable to define (Slaughter et al., 2025). One useful consequence of this is that credibility is not a design constant, but a design variable that platform designers and fact-checkers need to address as such. Transparency and procedural fairness need to be prioritized, and audiences subject to identity filters should be expected to interpret the same cues using these filters.

The second implication is the trade-off between reach and depth. Low-effort types of formats (labels) are more prone to appear in a feed environment and may typically yield a partial update result. In contrast, higher-effort explainers are better-corrected but are structurally at a disadvantage due to low click-through rates and low attention span (Okuhara et al., 2025). This synthesis thus assumes the general truths based on the fact that the most informative correction is the most preferred. The optimal correction is often that which is actually implemented in most platform situations. This is consistent with modeling evidence that the use of accuracy cues can decrease the sharing of misinformation because ensuring attention is diverted at the point of decision, even with no huge change at the time of decision making and belief (Lin et al., 2023). The policy implication of this distinction is that platforms can rightly boast of success in terms of minimizing spread, but the public-health or democratic disadvantages can still be experienced in case beliefs do not change.

The results further explain how format effectiveness is threatened by identity as well as reactance to form boundary conditions. Hard labels and removals are powerful interventions that can serve as declarations of deterrence, yet they might provoke censorship frames and draw attention towards other institutional goals than accuracy (Ecker et al., 2022). This does not imply that soft approaches are all the better. Rather, it indicates that strong formats are more varied interventions and their impact would be influenced by the trust of the intervening institution, perceived consistency of enforcement, and whether the users feel the correction is humiliating or face threatening consequences. The evidence that narrative corrections are not invariably superior to non-narrative corrections when informativeness is held constant substantiates the fact that avoiding threats is not merely a challenge of storytelling, but rather one of crafting explanations that are coherent and dignified yet offer a plausible alternative explanation (Ecker et al., 2020). This reflects a conditional design principle: formats ought to be designed where there is a high level of identity stake so that there is a reduction in the perceived coercive power and as many face-saving routes to revision.

The nature of platform conversational dynamics additionally makes the design of corrections difficult since the updating of beliefs is carried out within social environments, with audiences, rewards, and reputation risks. When applied publicly in threads, correction educates bystanders, and this is through social proof; however, when performative conflict occurs, it fuels polarization instead. The examples of WhatsApp and Facebook demonstrate that relational norms and cross-cutting exposure influence the attempt to act and reduce dysfunctional sharing with attempts to lower the cross-cutting exposure (Rossini et al., 2020). This underscores a weakness of most experimental studies that makes corrections to test as isolated messages. These kinds of designs tend to overlook the social costs of accepting error, incentives to indicate loyalty, and how corrections themselves might unintentionally reinforce misinformation. To do future qualitative work, further consideration of the order of interaction and who watches could be useful to explain the failure of technically correct corrections in the wild.

Lastly, time/visibility limitations highlight the fact that correction forms are processed in attention economies. The effectiveness of prebunking and media literacy intervention seems to be so in part due to their upstream nature, taking place even earlier on interpretation to close familiarity and motivated reasoning lock in (Roozenbeek et al., 2018; Basol et al., 2020; Guess et al., 2020). Nevertheless, these interventions continue to be challenged on the durability and topic generalization. The downstream debunking is not eliminated, but it is believed that it can work best when it is used to justify replacement as opposed to negating it completely and when it is introduced early enough to have an impact (Connor Desai and Reimers, 2023; Slaughter et al., 2025). The synthesis as a whole suggests that the discipline ought to cease thinking in terms of one best correction and begin conditional modeling that adapts correction forms to platform affordances, identity circumstances, and the feasibility of engagement with the user.

Conclusion

This article claimed that the forms of correction influence the belief updating not only through providing the correct information, but also organizing the social and cognitive environment in which individuals are willing and able to change what they believe to be true. On platforms, format served as a signal package and a set of constraints influencing credibility decision, focus allocation, identity threat, and replacement thinking opportunities. Corrections in which both credibility became visible by enabling transparency and demonstrating the source of knowledge may go further and validate acceptance.

Still, the same fact may be responded to with a counter-argument when users treat the corrections as partisan combat instead of epistemic remedies. Less effortful forms like labels and prompts were more likely to be experienced on fast-moving feeds, yet tended to result in partial updating, with less sharing on the part of the user that can be sustained and leads to a change in the belief. The explainers and narrative corrections of higher effort provided greater understanding and greater opportunities to substitute the misinformation in memory. Still, they had a disadvantage in terms of low adoption and platform conditions that encourage speed and emotionality.

Boundary conditions are also brought to the fore in the findings. The moderate forms with high levels of moderation might discourage propagation. Still, they may lead to legitimacy complaints due to the individual's feeling of control, whereas the community-based correction can cause less propagation but cannot avoid validity challenges. The timing was significant since interpretation could be reframed with prebunking and other upstream interventions prior to familiarity concretizing, but debunking demands alternate explanations of knowledge coherent enough, and visible enough, to rival exposure to misinformation repeated times. Generally, the research confirms one design logic, conditional, which holds that effective correction relies on a fit of format and platform affordances, as well as being corrected in a socially meaningful way. More direct testing on durability, more direct examination of contexts of personal messaging, and analysis of correction systems as dynamic sociotechnical systems instead of fixed types of messages should be part of future work.

References

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about Bad News: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*, *3*(1), 1–9. https://doi.org/10.5334/joc.91

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through the related stories functionality in social media. *Journal of Communication*, *65*(4), 619–638. https://doi.org/10.1111/jcom.12166

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *PNAS*, *114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Connor Desai, S., & Reimers, S. (2023). Does explaining the origins of misinformation improve the effectiveness of a given correction? *Memory & Cognition*, *51*, 422–436. https://doi.org/10.3758/s13421-022-01354-7

Ecker, U. K. H., Butler, L. H., & Hamby, A. (2020). You don't have to tell a story! A registered report testing the effectiveness of narrative versus non-narrative misinformation corrections. *Humanities and Social Sciences Communications*, *7*, 148. https://doi.org/10.1186/s41235-020-00266-x

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*, 13–29. https://doi.org/10.1038/s44159-021-00006-y

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993–1002. https://doi.org/10.1037/xge0000098

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *PNAS*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, *1*(1), 54–86. https://doi.org/10.1017/bpp.2016.2

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

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*(3), 106–131. https://doi.org/10.1177/1529100612451018

Lewin, S., Glenton, C., Munthe-Kaas, H., et al. (2015). Using qualitative evidence in decision making for health and social interventions: GRADE-CERQual. *PLOS Medicine*, *12*(10), e1001895. https://doi.org/10.1371/journal.pmed.1001895

Lewin, S., Glenton, C., Munthe-Kaas, H., et al. (2015). Using qualitative evidence in decision making for health and social interventions: GRADE-CERQual. *PLOS Medicine*, *12*(10), e1001895. https://doi.org/10.1371/journal.pmed.1001895

Lin, H., Pennycook, G., & Rand, D. G. (2023). Thinking more or thinking differently? Using drift-diffusion

modeling to illuminate why accuracy prompts decrease misinformation sharing. *Cognition, 230*, 105312. https://doi.org/10.1016/j.cognition.2022.105312

Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology, 54*, 101710. https://doi.org/10.1016/j.copsyc.2023.101710

Metzger, M. J., & Flanagin, A. J. (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics, 59*, 210–220. https://doi.org/10.1016/j.pragma.2013.07.012

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*, 303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*, 303–330. https://doi.org/10.1007/s11109-010-9112-2

Okuhara, T., Okada, H., Yokota, R., & Kiuchi, T. (2025). Effectiveness and determinants of narrative-based corrections for health misinformation: A systematic review. *Patient Education and Counseling, 139*, 109253. https://doi.org/10.1016/j.pec.2025.109253

Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2022). Shifting attention to accuracy can reduce misinformation online. *Nature Communications, 13*, 2463. https://doi.org/10.1038/s41467-022-30073-5

Roozenbeek, J., & van der Linden, S. (2018). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*. https://doi.org/10.1080/13669877.2018.1443491

Rossini, P., Stromer-Galley, J., Baptista, E. A., & de Oliveira, V. V. (2020). Dysfunctional information sharing on WhatsApp and Facebook: The role of political talk, cross-cutting exposure and social corrections. *New Media & Society, 23*(8), 2430–2451. https://doi.org/10.1177/1461444820928059

Slaughter, I., Peytavin, A., Ugander, J., & Saveski, M. (2025). Community notes reduce engagement with and diffusion of false information online. *PNAS, 122*(38), e2503413122. https://doi.org/10.1073/pnas.2503413122

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology, 8*, 45. https://doi.org/10.1186/1471-2288-8-45

van der Linden, S., Leiserowitz, A., Rosenthal, S., & Maibach, E. (2017). Inoculating the public against misinformation about climate change. *Global Challenges, 1*(2), 1600008. https://doi.org/10.1002/gch2.201600008

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation on social media. *Science Communication, 39*(5), 621–645. https://doi.org/10.1177/1075547017731776

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication, 37*(3), 350–375. https://doi.org/10.1080/10584609.2019.1668894

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior, 41*(1), 135–163. https://doi.org/10.1007/s11109-018-9443-y

Zhang, Y., et al. (2024). Correction by distraction: How high-tempo music enhances medical experts' debunking TikTok videos. *Journal of Computer-Mediated Communication, 29*(5), zmae007. https://doi.org/10.1093/jcmc/zmae007

Zhu, X., et al. (2025). Can correction messages reduce the spread of fake news on social media? *Journal of Management Information Systems*. https://doi.org/10.1080/07421222.2025.2520174