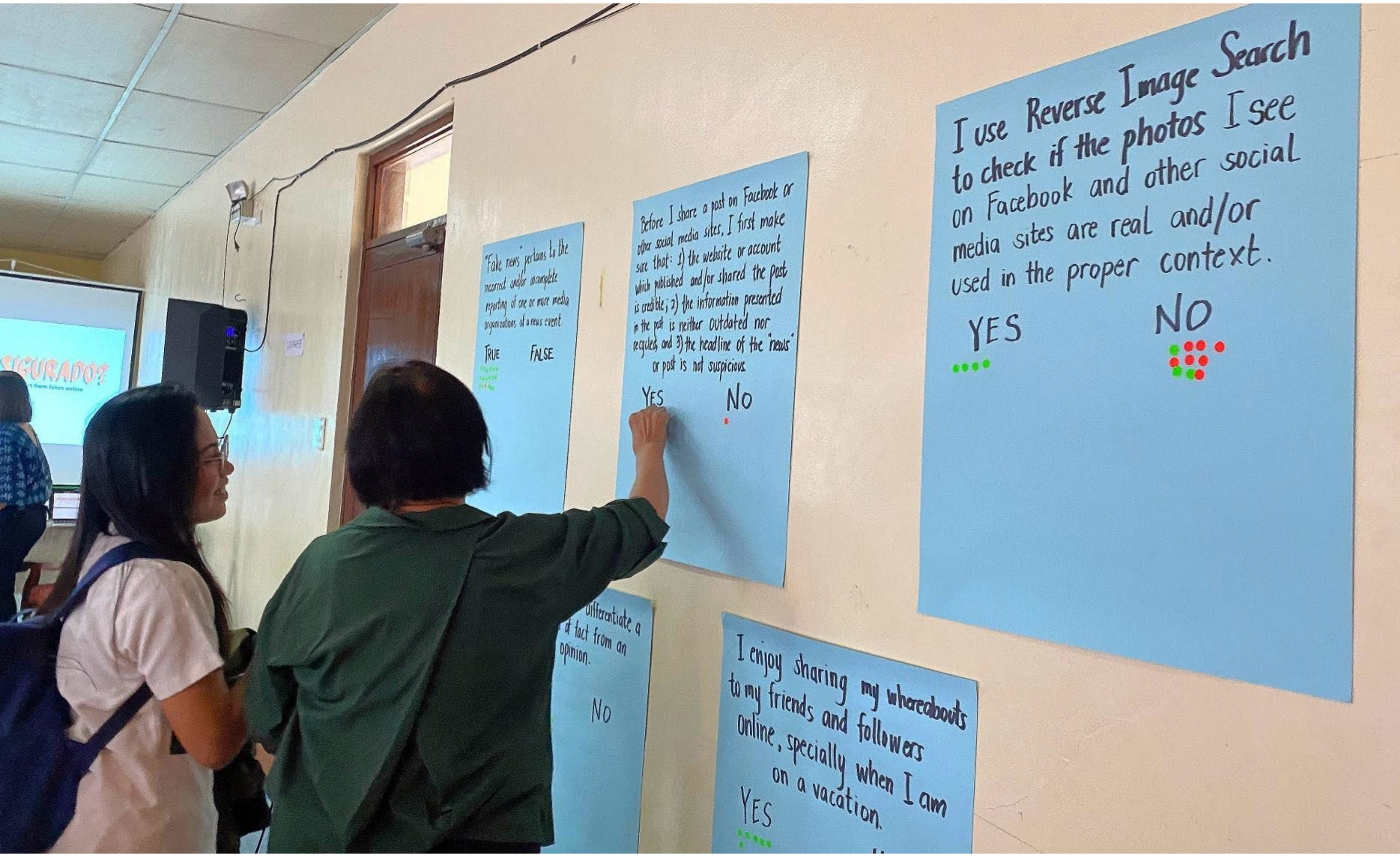
"Fake news" pertains to the incorrect and/or incomplete reporting of one or more media organizations of a news event
TRUE
FALSE
Before I share a post on Facebook or other social media sites, I first make sure that: 1) the website or account which published and/or shared the post is credible; 2) the information presented in the post is neither outdated nor recycled; and 3) the headline of the "news" or post is not suspicious.
YES
NO
I use Reverse Image Search to check if the photos I see on Facebook and other social media sites are real and/or used in the proper context.
YES
NO
differentiate a of fact from an opinion.
NO
I enjoy sharing my whereabouts to my friends and followers online, specially when I am on a vacation.
YES
SIGURADO?

USAID / PHILIPPINES

PHILIPPINES INFORMATION INTEGRITY EVIDENCE REVIEW

DISCLAIMER: The authors' views expressed in this publication do not necessarily reflect the views of the United States Agency for International Development or the United States Government.

CONTENTS

ACRONYMS

IMF	Initiative for Media Freedom
USAID	United States Agency for International Development

SUMMARY INFORMATION

This report provides an evidence-based foundation to understand information integrity interventions in the Philippines, both to assess the efficacy of current United States Agency for International Development (USAID)/Philippines programming and to help design future interventions. We build off existing reviews conducted by USAID of information integrity activities in the Global North by incorporating Philippines-specific contextual considerations and providing guidance to the Mission on what programmatic approaches are most likely to be effective in the Philippine context. The below information is structured in three sections. First, we outline the theory of change behind one of the two primary families of interventions to combat misinformation employed by the Initiative for Media Freedom (IMF) program (demand-side information integrity interventions) and highlight three contextual features of the misinformation environment in the Philippines that shape the applicability of the theory of change.[1] Second, we go through several specific types of interventions within the demand-side intervention family, detailing some broad conclusions about their effects from the existing literature, describing IMF programming in these categories supported by USAID, and discussing which conclusions best apply to the Philippines case. Third, we finish with recommendations about how to prioritize future programming and research.

OUTCOMES CONSIDERED	● **Sharing** and disseminating misinformation. ● **Belief** in misinformation. ● Strength of the **traditional media** sector.
PROGRAMMATIC APPROACH	● "Demand-side" interventions that reduce citizen susceptibility to misinformation. ● "Supply-side" interventions that strengthen media self-regulation and enhance journalistic integrity.
VARIANTS ON THE PROGRAMMATIC APPROACH CONSIDERED	● **Pre-information:** Inoculation, media literacy training, socio-psychological nudges. ● **Post-information:** Debunking ● **Supply side:** Journalist training, resources, and network-building.
STUDIES INCLUDED	● Experimental (30 studies) ● Quasi-experimental (1 study) ● Observational (11 studies)

[1] The IMF also employs "supply side" interventions that seek to improve the resiliency and capacity of media, but these interventions are not the focus of this evidence review.

LEVEL OF CONFIDENCE

Throughout the evidence review, individual conclusions include a small symbol like that below.

O	E

A shaded in E means the finding is based on experimental or quasi-experimental evidence (like a randomized controlled trial) and a shaded in O means the finding is based on (or also includes) observational evidence without a comparison group or estimate of the counterfactual. If just the bottom cells are shaded then the finding is based on one or a small number of studies, and if the top cells are shaded in, it means the finding is consistent across a number of studies.

CONTEXTUAL CONSIDERATIONS

Contextual factors that are pronounced for interpreting the application of studies to the Philippines:

- **Topic of misinformation:** Overtly political misinformation is common, relative to health and climate misinformation.
- **Traditional and digital literacy rates**: Traditional literacy is especially high in the Philippines, while digital literacy is very low.
- **Digital media environment:** The Philippines is characterized by a high prevalence of group-based social media, video-based misinformation, and political attacks on efforts to combat the spread of misinformation.

SUMMARY

"Demand-side" interventions meant to reduce citizen belief in and dissemination of misinformation are often effective, including in the Global South. High traditional literacy and low digital literacy rates in the Philippines provide an environment conducive to these interventions, though the political and digital media environments provide key challenges that limit the applicability of existing research. Future "demand-side" activities by the USAID-supported IMF project could be further enhanced by focusing on repeated citizen exposure to messaging campaigns, considering whether the dissemination strategies are reaching the most important target audiences, and catering messaging to reduce backlash effects. Less systematic evidence exists for "supply-side" interventions, though this is not to suggest that these programs are ineffective. Instead, expert surveys suggest these programs may be *more* effective in contexts like the Philippines. USAID would benefit from continuing to support these efforts, but should build in ways to define and measure the intended outcomes of this programming.

THEORY OF CHANGE

DEMAND-SIDE INTERVENTIONS

Demand-side interventions, which target citizen consumers, focus on reducing the likelihood that citizens 1) *believe* misinformation and 2) *share* misinformation with others in their network. The key distinction between the interventions discussed in this document is whether citizens are treated before (inoculation, media literacy training) or after (debunking) they are exposed to a specific piece of misinformation. When it comes to inoculation and media literacy training, the idea is to raise citizens' awareness of the prevalence of misinformation and the tactics used to spread it while arming them with techniques to discern between fact and fiction. Inoculation interventions often target specific types of misinformation in short, easily digestible messages and videos, while media literacy training (especially in Global South settings) tends to provide a more in-depth examination of the misinformation landscape, often in a group setting. The goal of debunking, or "fact-checking," is narrower and focuses on reducing the effects of particularly pernicious instances of misinformation after citizens are already exposed to it.

The USAID-supported IMF program engages in several demand-side interventions, including things categorized by the literature as "media literacy training." This entails in-person and Zoom workshops targeting students, youth leaders, online influencers, and a variety of other community members at the local level. It also includes shorter-form online media literacy training videos featuring prominent Filipino journalists and inoculation alerts and explainers posted on prominent social media sites about online disinformation and foreign influence operations. IMF also supports several debunking interventions, including a disinformation reporting platform and fact-checking websites, social media pages, videos, and alerts. IMF does not directly implement other demand-side interventions covered in the USAID report on countering misinformation in the Global South such as credibility tags and contextual labels.

SUPPLY-SIDE INTERVENTIONS

Supply-side interventions target the producers of information, with a focus on enhancing the capacity and quality of the traditional media sector. A variety of training programs and manuals aim to strengthen journalists' skills for producing high-quality, fact-checked information, with the goals of 1) reducing the spread of misinformation in the traditional media sector and 2) providing citizens with a viable alternative to less-reliable online information sources. Interventions sponsored by IMF that aim to achieve these goals include creating journalism associations and networking opportunities, implementing journalism safety guides and trainings, and providing capacity-building support for well-qualified journalists and organizations and legal support for vulnerable journalists. Excluded from this report are other "supply-side" interventions that are less feasible for USAID to support, such as social media platform alterations (e.g., encouraging people to leave Facebook) and alterations to politician messaging. It is also important to note that this report focuses specifically on the effect of supply-side interventions on outcomes related to misinformation; it does not speak to whether these interventions affect other important outcome categories.

PHILIPPINES CONTEXTUAL FACTORS THAT MODERATE THE THEORY OF CHANGE

In the following sections, this report considers lessons learned from the academic literature that portray the potential effectiveness of demand-side interventions to reduce the spread of misinformation in the Philippines. Several contextual features specific to the Philippines case, described in more detail below, are important for interpreting the applicability of the theory of change and findings in those studies. The effect of these contextual features on the interpretation of existing research is nuanced and varies by the specific program. However, the general takeaway is that efforts to combat information in the Philippines should be 1) ***harder*** as a result of the prevalence of overtly political misinformation, 2) ***easier*** as a result of the combination of high traditional literacy and low digital literacy, and 3) ***undetermined*** as a result of the digital media environment.

1. **Topics of misinformation: Overtly political misinformation in the Philippines.** One key aspect of the theory of change for many studies is that citizens are sufficiently open to new information and are willing to update their beliefs. This varies by the topic of misinformation, and studies on the spread of misinformation tend to focus on misinformation regarding 1) health (especially COVID-19), 2) climate change, and 3) politics. Although all three types of misinformation are prevalent in the Philippines, most of the misinformation is overtly political. For example, in 2022, 82 percent of reports to TotooBa.Info (a prominent fact-checking website established by IMF) were related to domestic politics and history (Fallorina et al, 2023). This includes recent misinformation cascades surrounding the Drug War, attacks against the press, the history of martial law, and Chinese incursions into the South China Sea (Rappler, “Patient Zero” report, funded by IMF). The body of existing research suggests that it is especially difficult to combat misinformation on topics that are tied to citizens’ political identities because “motivated reasoning” leads partisans to disregard efforts to combat misinformation that favors their preferred candidates or parties. Efforts to reduce the spread of political disinformation may also be difficult in settings like the Philippines, where there are organized attempts to discredit these campaigns (Ong and Cabañes, 2018).
2. **Literacy of target populations: High “traditional literacy” and low “digital literacy” in the Philippines.** The theory of change behind efforts to combat misinformation requires a certain traditional literacy threshold for target populations such that materials can be properly understood. The discrepancy between findings in the Global South and findings in the Global North is sometimes attributed to low traditional literacy rates (Blair et al., 2023); for example, Guess et al. (2020) found that the same intervention that worked among a highly-literate online population in India did not work in a rural setting with low literacy rates. At the same time, the theory behind media literacy trainings requires a low baseline level of “digital literacy,” a concept that captures whether people 1) think critically about online information, 2) know how to fact-check sources and acquire accurate information, 3) understand common sources of disinformation, 4) know how to protect their online privacy, and 5) feel confident engaging with news. In the Philippines, the traditional literacy rate is 98 percent, among the highest in the Global South (United Nations Educational Scientific and Cultural Organization Institute for Statistics), but the digital literacy rate in the Philippines, especially among younger citizens, is among the lowest of any country in the world (Association of Southeast Asian Nations Digital Literacy Programme).[2] Together, this suggests that target populations in the Philippines, especially students and youth leaders, may be especially ripe for interventions that counter misinformation.

[2] The Association of Southeast Asian Nations Digital Literacy Programme can be found at: https://assets.nationbuilder.com/aseanfoundationorg/pages/937/attachments/original/1710921777/ASEAN_DLP_Research_Report_-_One_Divide_or_Many_Divides.pdf?1710921777#page=92.39. A detailed description can be found at: https://btf.rappler.com/257/asean-foundation-unveils-research-findings-on-digital-literacy-spotighting-the-digital-divide-across-the-region/

3. **Digital media environment: Prevalence of misinformation spread in group settings and through video-based social media in the Philippines.** Two features of the digital media environment in the Philippines are important for interpreting the effects of efforts to combat misinformation. First, the theory of change for several interventions is based on "social identity theory," which suggests that efforts to combat misinformation are more effective when they come from "in-group" members, whether those be members of the same social group (family, friends), identity group (ethnic, religious), or political group (party, organization). Concurrently, this theory implies that organized attempts to undermine efforts to combat disinformation can be intensified by group dynamics. In the Philippines, disinformation spread is particularly prevalent in "group" settings, including Facebook groups, YouTube channels, and group chats of family and friends on WhatsApp, Messenger, or Telegram (Ong & Cabañes, 2018). This can be both a blessing and a curse, as group dynamics can enhance the spread of disinformation but may also be harnessed in efforts to combat disinformation. Second, existing theory tends to focus on combating misinformation delivered through text- and image-based sources. Although this theory should also apply to misinformation delivered via video in principle, it is possible that this medium activates different psychological mechanisms. Recently, the Philippines has seen an increase in misinformation on TikTok and YouTube, and the evidence is very sparse when it comes to combating misinformation on these mediums.

INTERVENTION EFFICACY IN THE PHILIPPINES

Detailed evidence exists on the causal effect of several demand-side interventions, though studies on whether these insights apply to the Global South are new. On the other hand, reliable evidence on the effects of supply-side interventions is severely limited. This section describes the main findings on the interventions covered in the Blair et al. (2023) USAID report, outlines existing activities supported by IMF in each category, and discusses the applicability of the literature to the Philippines setting.

INOCULATION AND MEDIA LITERACY (PRE-INFORMATION)

The first category of interventions attempts to reduce citizen susceptibility before they are exposed to a specific piece of misinformation. Inoculation interventions provide short messages warning citizens about common techniques to misinform people, supply examples of false stories, and let citizens practice identifying misinformation. Media literacy trainings build citizens' skills for identifying common types of misinformation and verifying the accuracy of information. The "trainings" evaluated in the literature range from short infographics and videos to multi-week courses on digital literacy. Socio-psychological nudges (including accuracy prompts, friction prompts, and social norm prompts) are short messages that urge readers to slow down and think about the accuracy of information before reading it.

MAIN LITERATURE TAKEAWAYS O E

- Inoculation tends to be effective at reducing citizen susceptibility to misinformation, including in the Global South. Socio-psychological nudges produce similar results, though effect sizes are small and more experimental work is needed (Blair et al., 2023).
- The effects of media literacy training are mixed. These interventions produce largely null results in the Global North, though the "trainings" tested in these settings are often very short-form infographics, articles, or short (30-second) videos. In the Global South, media literacy trainings tend to produce positive effects among highly educated populations

(Guess et al., 2020) and when they appeal to users' emotions and past engagement with false news (Ali and Qazi, 2021; Athey et al., 2022; Gottlieb et al., 2022). Long-term media literacy training targeted at university students also find positive effects (Apuke et al., 2022, 2023; Zhang et al., 2022).

APPLICATIONS TO IMF INTERVENTIONS AND THE PHILIPPINES CONTEXT

The IMF supports a wide range of programs in this space. In particular, media literacy trainings (both online and in-person) are the centerpiece of demand-side activities supported by IMF:

- **Long-form workshops, webinars, video tutorials, bootcamps, and roadshows that teach fact-checking and digital literacy skills.**
- **Short-form infographics, comic strips, videos featuring prominent journalists, radio programs, and gaming applications.**
- **Materials are disseminated through a Facebook page (BarangayHub) and partnerships with existing media organizations and individual influencers.**

Comparisons between the **interventions** typically studied in the literature and those implemented by IMF have important implications for the applicability of previous findings:

Long-form training worked among university students but generated backlash among regime supporters. Only a small group of studies (all of which were implemented in the Global South) examines media literacy trainings that are as in-depth as the long-form trainings implemented by IMF. Apuke et al, (2022, 2023) and Zhang et al. (2022) evaluated an eight-week course on digital and media literacy at a university in Nigeria, finding that the participating students became less likely to believe and share misinformation. Badrinathan (2021) implemented an hour-long training on media literacy skills in India, in which enumerators delivered the training directly to individual subjects at their homes. The treatment produced null effects overall and generated a backlash effect among Bharatiya Janata Party supporters, who became *more* likely to believe misinformation as a result of the training. The latter study calls into question the effect that IMF programming may have among Marcos or Duterte supporters, given that these politicians often attack efforts to combat misinformation as being biased against them.

Short-form media literacy training produced largely null effects in the Global North. In the Global South, they reduced susceptibility to misinformation only when they 1) were accompanied by personalized feedback on users' past engagement and 2) appealed to users' emotions rather than focusing simply on teaching media literacy skills. A larger group of studies examines the effect of short-form media literacy training (videos, infographics, articles) similar to those implemented by IMF. Studies in the Global North find that presenting media literacy education through short articles and pop-ups (Hameleers, 2022; Vraga et al., 2022; Panizza et al., 2022), infographics (Qian et al., 2023; Domgaard & Park, 2021) and videos (Vraga et al., 2021) all produced null effects on subjects' belief in misinformation. Aslett et al. (2022) found a *negative* effect of encouraging users to utilize search engines to verify information. Although the effects of standard short-form media literacy interventions in the Global South were also largely null, studies find that short (three- to four-minute) media literacy training videos are effective when paired with personalized feedback on users' ability to identify misinformation (Ali and Qazi, 2021; Bowles et al., 2023) and when the videos tried to elicit empathy for outgroup members (Gottlieb et al., 2022).

Repeated delivery of information to the same individuals over time is more effective than isolated training. The short-form media literacy and inoculation interventions supported by IMF are less systematic about following up with users compared to the most effective interventions tested in the literature. Existing efforts include directing participants to the BarangayHub Facebook page, but this “passive” form of follow-up seems less effective than “active” follow-up. For example, Bowles et al. (2023) found that incentivizing participants to participate in regular fact-checking quizzes significantly increased the effects of fact-checks sent via WhatsApp. The studies that found the most consistent positive effects of inoculation and media literacy training (especially in Global South settings) tend to consistently follow up with participants via emails, SMS/WhatsApp messages, or mobile app notifications. The effects of one-time prompts and training are more mixed and less likely to have lasting effects. For example, the three field experiments conducted in the Global South found consistently positive effects of inoculation and debunking that primed citizens with information every few weeks for a period of several months (Bowles et al., 2023; Pereira et al., 2022; Garg et al., 2022). Garg et al. (2022) found that it took nine weeks of repeated inoculation messages before citizens’ truth discernment significantly improved. On the other hand, the field experiments that implemented one-time in-person media literacy training (Badrinathan, 2021) and online inoculation exercises (Iyengar et al., 2022; Ma et al., 2023) found null effects, or effects that lasted no more than one week. Even the most promising study showing a positive effect of a short-form media literacy intervention found that results diminished substantially after a couple of weeks (Guess et al., 2020).

Simple, short messages can be just as effective as more in-depth training. Although the effects of both long-form and short-form media literacy training are mixed, studies often find that short “accuracy nudges” imploring users to be skeptical of headlines and consider their accuracy have more consistently positive effects. In the most direct comparison of different delivery methods, Bowles et al. (2023) found that short, text-based inoculation messages sent via WhatsApp were more effective at reducing susceptibility to misinformation than four- to eight-minute podcasts. Offer-Westort et al. (2024) found positive effects of short accuracy nudges delivered via a Facebook Messenger chatbot, and Athey et al. (2022) found positive effects of short accuracy nudge text messages sent over five consecutive days. Most notably, the only experimental study in the Blair et al. (2023) review conducted in the Philippines (Arechar et al., 2023) found positive effects of these shorter-form interventions.

DEBUNKING (POST-INFORMATION)

MAIN LITERATURE TAKEAWAYS O E

- Debunking is generally effective at correcting misinformed beliefs, including in the Global South. However, the effects on behavioral change are small and inconsistent. There is little evidence that corrections backfire, especially when provided by expert sources. Corrections from sources that share personal, political, or religious ties with recipients tend to be more effective (Blair et al., 2023).

APPLICATIONS TO IMF INTERVENTIONS AND THE PHILIPPINES CONTEXT O E

Some of the more prominent examples of IMF debunking programs include:

- **Disinformation reporting websites, fact-checking videos, infographics, and social media posts.**
- **Dissemination of debunking materials through media partners, civil society organizations, and independent influencers.**

A review of the literature reveals several key findings that are relevant for IMF programming and future potential programming.

The majority of evidence on debunking is specific to 1) health (especially COVID-19) and 2) climate change. Less evidence exists relating to debunking overtly political misinformation in a competitive digital media environment. Especially in the Global South, studies tend to focus on debunking misinformation about public health (Carey et al., 2020; Porter et al., 2023; Porter and Wood, 2021; Bowles et al., 2020; Winters et al., 2021; Armand et al., 2021). The one study that expressly focused on political misinformation (Pereira et al., 2022) dealt with misinformation during the 2018 Brazilian election and found null results. Given the prevalence of political misinformation in the Philippines, implementers should be cognizant of the fact that debunking efforts are intentionally discredited by prominent politicians, potentially undermining their effects. None of the studies evaluate the effects of debunking in a "competitive" environment where users are also presented with counter-narratives that attempt to undermine fact-checking activities. Moreover, studies tend to find that debunking is more effective when it comes from "expert" sources that are perceived as credible and neutral (van der Meer and Jin, 2020; Vraga and Bode, 2017). Efforts to debunk political disinformation in the Philippines should be aware of how users view the political slant of particular partner news outlets. Debunking efforts should also take into account the credibility of the United States, especially after the recent revelations about Department of Defense misinformation campaigns surrounding Chinese COVID-19 vaccines.

Debunking efforts are more effective when corrections are endorsed by individuals or groups that share an identity or social tie with the consumer. Studies in India (Armand et al., 2021) and Pakistan (Pasquetto et al., 2022) found that consumers are more prone to believe debunking materials when they are endorsed by "in-group" members, such as family, members of shared identity groups, or members of the same political party. Given the group-based nature of the digital media environment in the Philippines and the importance of social ties, it is crucial to employ dissemination strategies that increase the likelihood of individuals receiving fact-checking information from others with whom they share social ties.

JOURNALIST CAPACITY-BUILDING (SUPPLY-SIDE)

MAIN LITERATURE TAKEAWAYS O E

- As Schiffrin and Berman state, "In spite of seemingly large amounts spent on journalism training, little publicly available literature shows how to evaluate journalism training courses and what to look for when assessing improvements" (2011, p. 341). More evidence is needed to verify 1) whether journalistic interventions improve the quality of traditional journalism and 2) whether improved journalism quality makes citizens less susceptible to misinformation.

APPLICATIONS TO IMF INTERVENTIONS AND THE PHILIPPINES CONTEXT

Support for the journalism sector is the second main centerpiece of IMF programming, along with media literacy trainings. Prominent examples include:

- **Journalism capacity development, including training, ethics guides, and support for journalist associations and media outlets.**
- **Legal protection, peer-to-peer networks, hotlines, and trainings to improve journalists' safety.**

There is a lack of high-quality evidence explicitly connecting journalist interventions to misinformation outcomes. The sole high-quality study that focuses on how interventions targeting journalists affect the spread of misinformation is Graves et al. (2016), who found that letters sent to journalists in the US appealing to journalistic norms are more effective at instigating fact-checking than letters focusing on readers' demands. More peripherally related, Michelitch and Weghorst (2021) evaluated a journalist training program in Tanzania, finding that their intervention did not improve journalists' knowledge, ethics, or gender diversity. When it comes to the second step in the causal chain (connecting journalist quality to citizen susceptibility to misinformation), high-quality evidence is also scarce. The best evidence available are studies that suggest data visualizations in news articles did not improve citizens' misperceptions about immigrants and COVID-19 vaccines (Mena, 2023) and news articles "forewarning" citizens about multiple sides of the argument on climate change reduced citizens' belief in the positions of science deniers.

Despite the lack of evidence, experts on misinformation in the Global South suggest allocating significant resources to journalist training. In their report on misinformation in the Global South, Blair et al. (2023) surveyed 138 experts on misinformation, including 89 researchers and 47 practitioners. The core question of the survey asked the experts how they would allocate resources to various interventions to counter misinformation. On average, the experts on the Global South suggested they would allocate over 15 percent of their resources to journalist trainings, tied with media literacy trainings as the most promising intervention category. Practitioners were especially optimistic about journalist trainings, allocating almost 20 percent of their hypothetical resources to this intervention, making it the top-rated intervention. This finding from the expert survey provides striking support for the need for more research on these types of interventions.

PHILIPPINES-SPECIFIC CONTEXT CAVEATS

Several features of the **Philippines context** shape how previous results should be interpreted:

- **Media literacy interventions are more likely to work when “traditional literacy” rates are already high.** As discussed, traditional literacy rates in the Philippines are high, including in rural areas. This has positive implications for the potential effectiveness of media literacy training, for users are more likely to properly digest the material. Most directly, Guess et al. (2020) found that the same media literacy intervention had positive effects among an educated online sample in India but no effect among a representative rural sample in the same country, consistent with the results in Badrinathan (2021). Second, the Arechar et al. (2023) study attributed their inconsistent findings of accuracy nudges across the Global South to lower education levels. It is notable that they found consistent positive effects in the Philippines in particular. In light of these facts, Blair et al.’s (2023) statement that “a highly educated sample may benefit from short, information-based interventions similar to their counterparts in the Global North” (p. 35) may be applicable to the context in the Philippines. This is especially the case for the well-educated and highly literate populations that often attend the IMF media literacy webinars and in-person trainings.
- **Media literacy training can backfire among supporters of political parties who view the efforts as being partisan.** The topic of combating misinformation is highly politicized in the Philippines, and much of the most pernicious misinformation is overtly political. However, the majority of demand-side interventions that reduce citizens’ susceptibility to misinformation focus on the topics of climate change and health, with a smaller number of studies that focus on overtly politicized topics (Arechar et al., 2023; Bowles et al., 2023). The limited number of studies in the Global South that focus on overtly political misinformation tend to find divergent effects based on the ideology of participants. Most prominently, Badrinathan (2021) finds that Bharatiya Janata Party supporters in India became *more* susceptible to misinformation after an intensive media literacy training. India represents one of the closest misinformation environments to the Philippines, increasing the salience that the risk of a similar backlash effect may be particularly high among Marcos or Duterte supporters.
- **Interventions to combat misinformation are more effective when they leverage social norms.** The digital media environment and misinformation landscape in the Philippines are both typified by high levels of group-based communication channels. Family message groups and Facebook groups are common mediums for the spread of misinformation, which is a blessing and a curse for efforts to combat misinformation. On the one hand, when users attempt to correct misinformation by group members in the Philippines, they are commonly attacked by disagreeing family members or armies of trolls (Ong and Cabanes, 2018). On the other hand, interventions that leverage social pressure and social norms tend to find more consistently positive effects in the literature (Gottlieb et al., 2022; Badrinathan and Chauchard, 2023; Bowles et al., 2023). This tradeoff is reflected in Pasquetto et al. (2020), who found that correcting misinformation is most effective when it comes from peers; however, this practice was unpleasant for the individuals correcting misinformation.
- **Existing evidence has little to say about how to combat video-based misinformation.** The Philippines has seen a recent rise in misinformation spread through

YouTube channels and TikTok. Existing studies have little to say about whether existing interventions are effective at combating misinformation spread through these mediums. Most studies measure the likelihood of believing and spreading misinformation by presenting participants with a list of text-based articles and asking them to assess their accuracy. Although the mechanism for combatting video-based misinformation may be the same, there is little direct evidence about how media literacy training operates in this space.

RECOMMENDATIONS

The above analysis of the literature suggests several recommendations for future IMF programming in the Philippines, starting with demand-side interventions and concluding with supply-side interventions.

1. **Invest in efforts to repeatedly follow up with participants in media literacy training.** IMF supports various activities that the literature would characterize as "media literacy training" that have been shown to be effective in certain contexts. However, the effects of one-time training are mixed, while inoculation efforts that repeatedly follow up with participants tend to show more positive results. For example, drawing from examples in the literature, implementers could follow up with regular messages to participants via social messaging platforms or by providing participants with subscriptions to trusted media partners. Other lessons from the literature on media literacy training suggest that the programs should continue to focus on recruiting more highly educated populations and appeal to them using emotions rather than informational lectures.
2. **Build on existing dissemination strategies to enhance the reach of inoculation and debunking interventions.** The literature suggests that the short-form inoculation and debunking videos and explainers supported by the IMF should be effective at reducing the spread of information, including among individuals who might be more conspiracy-minded. As a result, resources should be devoted to getting these materials in front of as many eyes as possible and to avoid "preaching to the choir" of individuals who are less likely to believe and spread misinformation. Some of these efforts are already being undertaken, such as partnering with influencers on social media sites. One key takeaway from the literature is that simplicity in messaging is a virtue, especially if it enhances efforts to improve the reach of interventions. To improve cost-effectiveness, greater investment could be focused on dissemination strategies (drawing from marketing research) rather than long-form interventions with more limited reach.
3. **Cater counter-disinformation campaigns to specific subgroups.** Evidence suggests that demand-side interventions to reduce the spread of information are more effective when shared by trusted sources but can create a backlash effect when they appear ideologically motivated. This may be especially true in the Philippines context, given the importance of 1) Facebook groups and pages, 2) group message platforms (Messenger, WhatsApp), and 3) troll farms to the disinformation landscape. Since misinformation campaigns are likely to be targeted by groups trying to undermine these efforts, it is crucial to understand how ideological lenses (both of the deliverer and recipient of misinformation campaigns) affect how these interventions are received. For example, before rolling out large-scale campaigns, it is crucial to conduct pilot tests to see whether interventions generate backlash.

4. **Specify the outcomes of interest and increase investment in monitoring and evaluation for supply-side interventions.** The theory of change for the supply-side interventions could be made more explicit, as could the outcomes that the interventions are trying to affect. Given the scarcity of evidence on supply-side interventions and the fact that experts think these interventions show great promise, it is worth continuing to invest in this programming and also invest in evaluating its effects on the specified outcomes. There are three core intended outcomes of this programming: 1) improving journalists' physical, legal, and socio-psychological safety; 2) improving the volume and quality of independent media; and 3) increasing citizens' reliance on high-quality independent media (rather than less reliable online information). The link between the interventions and the first two outcomes is relatively clear, but more work could be done to connect programming to citizen-facing outcomes.

REFERENCES

Ali, A., & Qazi, I. A. (2023). Countering misinformation on social media through educational interventions: Evidence from a randomized experiment in Pakistan. *Journal of Development Economics*, *163*, 103108.

Apuke, O. D., Omar, B., Tunca, E. A., & Gever, C. V. (2022). Information overload and misinformation sharing behaviour of social media users: Testing the moderating role of cognitive ability. *Journal of Information Science*, 01655515221121942.

Arechar, A. A., Allen, J., Berinsky, A. J., Cole, R., Epstein, Z., Garimella, K., ... & Rand, D. G. (2023). Understanding and combatting misinformation across 16 countries on six continents. *Nature Human Behaviour*, *7*(9), 1502–1513.

Armand, A., Augsburg, B., Bancalari, A., & Kameshwara, K. K. (2021). *Countering misinformation with targeted messages: Experimental evidence using mobile phones* (No. W21/27). Institute for Fiscal Studies.

Armand, A., Augsburg, B., Bancalari, A., & Kameshwara, K. K. (2021). *Social Proximity and Misinformation: Experimental Evidence from a Mobile Phone-Based Campaign in India* (No. 16492). CEPR Discussion Papers.

Armand, A., Augsburg, B., Bancalari, A., & Kameshwara, K. K. (2024). Religious proximity and misinformation: Experimental evidence from a mobile phone-based campaign in India. *Journal of Health Economics*, *96*, 102883.

Aslett, K., Guess, A. M., Bonneau, R., Nagler, J., & Tucker, J. A. (2022). News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. *Science advances*, *8*(18), eabl3844.

Athey, S., Cersosimo, M., Koutout, K., & Li, Z. (2022). Emotion-versus reasoning-based drivers of misinformation sharing: A field experiment using text message courses in Kenya.

Badrinathan, S. (2021). Educative interventions to combat misinformation: Evidence from a field experiment in India. *American Political Science Review*, *115*(4), 1325–1341.

Badrinathan, S., & Chauchard, S. (2023). Researching and countering misinformation in the Global South. *Current Opinion in Psychology*, 101733.

Blair, R. A., Gottlieb, J., Nyhan, B., Paler, L., Argote, P., & Stainfield, C. J. (2023). Interventions to counter misinformation: Lessons from the Global North and applications to the Global South. *Current Opinion in Psychology*, 101732.

Bowles, J., Larreguy, H., & Liu, S. (2020). Countering misinformation via WhatsApp: Preliminary evidence from the COVID-19 pandemic in Zimbabwe. *PloS one*, *15*(10), e0240005.

Bowles, J., Croke, K., Larreguy, H., Marshall, J., & Liu, S. (2023). Sustaining exposure to fact-checks: Misinformation discernment, media consumption, and its political implications. *Media Consumption, and its Political Implications (September 25, 2023)*.

Carey, J. M., Chi, V., Flynn, D. J., Nyhan, B., & Zeitzoff, T. (2020). The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil. *Science advances*, *6*(5), eaaw7449.Carey, J. M., Chi, V., Flynn, D. J., Nyhan, B., & Zeitzoff, T. (2020). The effects of corrective information about disease epidemics and outbreaks: Evidence from Zika and yellow fever in Brazil. *Science advances*, *6*(5), eaaw7449.

Domgaard, S., & Park, M. (2021). Combating misinformation: The effects of infographics in verifying false vaccine news. *Health Education Journal*, *80*(8), 974–986.

Fallorina, R., Lanuza, J. M. H., Felix, J. G., Sanchez II, F., Ong, J. C., & Curato, N. (2023). The Evolution of Disinformation in Three Electoral Cycles.

Gisondi, M. A., Barber, R., Faust, J. S., Raja, A., Strehlow, M. C., Westafer, L. M., & Gottlieb, M. (2022). A deadly infodemic: social media and the power of COVID-19 misinformation. *Journal of medical Internet research*, *24*(2), e35552.

Graves, L., Nyhan, B., & Reifler, J. (2016). Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. *Journal of Communication*, *66*(1), 102–138.

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. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545.

Hameleers, M. (2022). Separating truth from lies: Comparing the effects of news media literacy interventions and fact-checkers in response to political misinformation in the US and Netherlands. *Information, Communication & Society*, *25*(1), 110–126.

Iyengar, A., Gupta, P., & Priya, N. (2023). Inoculation against conspiracy theories: A consumer side approach to India's fake news problem. *Applied Cognitive Psychology*, *37*(2), 290–303.

Kim, S. C., Vraga, E. K., & Cook, J. (2021). An eye tracking approach to understanding misinformation and correction strategies on social media: The mediating role of attention and credibility to reduce HPV vaccine misperceptions. *Health Communication*, *36*(13), 1687–1696.

Kozyreva, A., Lorenz-Spreen, P., Herzog, S., Ecker, U., Lewandowsky, S., Hertwig, R., ... & Wineburg, S. (2022). Toolbox of interventions against online misinformation and manipulation.

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

Mena, P. (2023). Reducing misperceptions through news stories with data visualization: The role of readers' prior knowledge and prior beliefs. *Journalism*, *24*(4), 729–748.

Michelitch, K., & Weghorst, K. (2021). *Impact evaluation of an intensive journalism training activity in Tanzania: Final report* (Contract No. GS-10F-0033M / AID-OAA-M-13-00013, Tasking N061). NORC at the University of Chicago.

Offer-Westort, M., Rosenzweig, L. R., & Athey, S. (2024). Battling the coronavirus 'infodemic'among social media users in Kenya and Nigeria. *Nature Human Behaviour*, *8*(5), 823–834.

Ong, J. C., & Cabañes, J. V. A. (2018). Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines. *Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the Philippines*.

Pasquetto, I. V., Jahani, E., Atreja, S., & Baum, M. (2022). Social debunking of misinformation on WhatsApp: the case for strong and in-group ties. *Proceedings of the ACM on human-computer interaction*, *6*(CSCW1), 1-35.

Pereira, F., Bueno, N. S., Nunes, F., & Pavão, N. (2022). Fake news, fact checking, and partisanship: the resilience of rumors in the 2018 Brazilian elections. *The Journal of Politics*, *84*(4), 2188–2201.

Pereira, F. B., Bueno, N. S., Nunes, F., & Pavão, N. (2023). Inoculation Reduces Misinformation: Experimental Evidence from Multidimensional Interventions in Brazil. *Journal of Experimental Political Science*, 1–12.

Porter, E., & Wood, T. J. (2021). The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom. *Proceedings of the National Academy of Sciences*, *118*(37), e2104235118.

Schiffrin, A., & Behrman, M. (2011). Does training make a difference? Evaluating journalism training programs in Sub-Saharan Africa. *Journalism & Mass Communication Educator*, *66*(4), 340–360.

Sharma, D. K., Shrivastava, P., & Garg, S. (2022, March). Utilizing word embedding and linguistic features for fake news detection. In *2022 9th International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 844–848). IEEE.

Van der Meer, T. G., & Jin, Y. (2020). Seeking formula for misinformation treatment in public health crises: The effects of corrective information type and source. *Health communication*, *35*(5), 560–575.

Velez, Y. R., Porter, E., & Wood, T. J. (2023). Latino-targeted misinformation and the power of factual corrections. *The Journal of Politics*, *85*(2), 789–794.

Vraga, E. K., & Bode, L. (2017). Using expert sources to correct health misinformation in social media. *Science communication*, *39*(5), 621–645.

Vraga, E. K., Bode, L., & Tully, M. (2022). Creating news literacy messages to enhance expert corrections of misinformation on Twitter. *Communication Research*, *49*(2), 245–267.

Wei, L., Gong, J., Xu, J., Abidin, N. E. Z., & Apuke, O. D. (2023). Do social media literacy skills help in combating fake news spread? Modelling the moderating role of social media literacy skills in the relationship between rational choice factors and fake news sharing behaviour. *Telematics and Informatics*, *76*, 101910.

Winters, M., Oppenheim, B., Sengeh, P., Jalloh, M. B., Webber, N., Pratt, S. A., ... & Nordenstedt, H. (2021). Debunking highly prevalent health misinformation using audio dramas delivered by WhatsApp: evidence from a randomised controlled trial in Sierra Leone. *BMJ global health*, *6*(11), e006954.

Zhang, L., Iyendo, T. O., Apuke, O. D., & Gever, C. V. (2022). Experimenting the effect of using visual multimedia intervention to inculcate social media literacy skills to tackle fake news. *Journal of Information Science*, 01655515221131797.

Zhou, Y., Yang, Y., Ying, Q., Qian, Z., & Zhang, X. (2023, June). Multi-modal fake news detection on social media via multi-grained information fusion. In *Proceedings of the 2023 ACM international conference on multimedia retrieval* (pp. 343–352).