Series Name: WPS
WPS Paper No.: 213
Issue Date: 9/19/23

The spread of (mis)information: A social media experiment in Pakistan

Sarojini Hirshleifer, Mustafa Naseem, Agha Ali Raza, and Arman Rezaee

CEGA

Center for Effective Global Action

Working Paper Series

Center for Effective Global Action

University of California

The spread of (mis)information: A social media experiment in Pakistan*

Sarojini Hirshleifer†, Mustafa Naseem‡, Agha Ali Raza§, and Arman Rezaee¶

September 19, 2023

Abstract

This study examines the dissemination of (mis)information on a social media platform in Pakistan. It combines an intervention to disseminate official information about the COVID-19 pandemic across the platform with a randomized experiment that measures the impact of fully controlling access to pandemic-related misinformation. The two treatments rely on a higher-intensity, *ex-ante* approach to moderating misinformation on the platform relative to the control, which relies on a more standard *ex-post* approach to moderation. In one treatment, no misinformation was allowed on the platform, while in the other, it was allowed with an official rebuttal. Controlling misinformation, as in the treatments, reduces platform usage by 41%, indicating a distaste for moderation. Furthermore, the treatments reduce exposure to official information by 29% more than they reduce exposure to misinformation. A conceptual framework posits that these findings can be explained by the fact that, in this setting, official information is more trusted, and thus is more widely disseminated, relative to misinformation. We find evidence for two potential mechanisms for the observed distaste for moderation.

Keywords: social media, misinformation, information, digital economy, political economy, development economics, health, COVID-19, field experiment, randomized controlled trial
JEL Codes: L86, L82, D80, D83, O10, I10, I15, P00

*We gratefully acknowledge that this work was partially supported by the National Institutes of Health grant 5R21HD095696-02. We thank Fizzah Malik, Namoos Hayat Qasmi, Shan Randawa, Sacha St-Onge Ahmad, and Behzad Taimur for providing support for the implementation of the project and for extracting the data from the platform. We also thank Marcella Alsan, Natalie Bau and seminar participants at the CEGA Research Retreat, University of Oregon, UC Irvine, BREAD, and NBER Summer Institute: Digital Economics and Artificial Intelligence for helpful feedback.

†University of California, Riverside and CEGA (email: sarojini.hirshleifer@ucr.edu)
‡University of Michigan (email: mnaseem@umich.edu)
§Lahore University of Management Sciences (email: agha.ali.raza@lums.edu.pk)
¶University of California, Davis and CEGA (email: abrezaee@ucdavis.edu)

1 Introduction

Social media is a powerful tool that can dramatically reduce the cost of information sharing and reach people who may not regularly engage with formal media. That any user can share content from any source, however, poses a risk: social media can allow both helpful, accurate information and harmful, inaccurate information to be disseminated much more widely than it otherwise would have. Both of these potential roles for social media are particularly relevant during times of crisis, such as political events, natural disasters, and the COVID-19 pandemic. On the one hand, sharing high-quality information on social media is widely recognized as an important tool for policymakers.[1] On the other hand, misinformation is especially likely to spread on social media, and is a critical risk factor in the harms caused by crises.[2] Although managing the dissemination of information on social media is a major policy challenge throughout the world, it is particularly relevant to developing countries where high-quality information is more likely to be scarce.[3]

Using a randomized experiment, we study the impact of fully controlling access to misinformation across a social media platform on users' exposure to both official, high-quality information and misinformation itself. In order to control access to misinformation, it must first be identified. This requires ex-ante moderation of *all* information on the platform. Thus, a complete approach to controlling misinformation also has implications for the dissemination of official information. In addition, conditional on fully controlling misinformation, a secondary question is how to address it. We examine two approaches: never exposing misinformation or rebutting it directly.

This study took place in the context of a social media platform in Pakistan called Baang, early in the COVID-19 pandemic. The study had two main components. First, we used the platform to disseminate information about COVID-19 in the form of *official* posts from Baang. This content was available to all users of the platform throughout the study. Second, we implemented a novel,

[1] Government and official use of social media is widespread, and is recommended by researchers given high levels of engagement from users (Lin et al., 2016; Tursunbayeva, Franco and Pagliari, 2017). For example, the crisis communication plan of the CDC highlights the value of disseminating information through such platforms (CDC, 2018). In addition, during the early days of the pandemic, social media companies used their platforms to actively disseminate relevant information. Facebook had a coronavirus information center at the top of news feeds for a time (Dwoskin, 2020), while Twitter has had a section on its Explore tab dedicated to news on COVID-19 (Twitter, 2020).

[2] Allcott and Gentzkow (2017) document the role of social media in spread of "fake news" during the 2016 U.S. presidential election. During the pandemic, the Director-General of the World Health Organization (WHO) noted in 2020, "...we're not just fighting an epidemic; we're fighting an infodemic. Misinformation has also been challenge in addressing natural disasters (Hsu, 2023). Fake news spreads faster and more easily than this virus, and is just as dangerous" (WHO, 2020). More generally, health misinformation on social media has been a concern since before the pandemic (Wang et al., 2019).

[3] U.S. government agencies have frequently interacted directly with social media companies, which highlights the relevance of the dissemination of information on social media to first-order policy issues. The limits of this interaction a topic of ongoing debate (Fung and Cole, 2023). Developing countries are widely perceived to be information scarce, particularly with regards to health information (Stiglitz, 2000; Kremer and Glennerster, 2011).

user-level randomized experiment that varied the approach to controlling user-generated misinformation about COVID-19 on the platform. In the control condition, users have access to a version of the platform that relies on lower-intensity *ex-post* moderation to address misinformation. This is conceptually aligned with a traditional approach to moderation on most social media platforms. The treatment versions of the platform, however, rely on higher-intensity *ex-ante* moderation: all user-generated content is reviewed by a moderator before being posted on the platform. In the *remove* treatment, content that includes misinformation is simply never posted to the platform, while in the *sunshine* treatment, misinformation is posted along with a rebuttal that debunks it. These rebuttals include high-quality information in line with official posts. A user's own condition assignment does not affect how their posts are distributed on the platform. Thus, aside from the posts that included misinformation, all content on the platform was the same in all of the conditions.

The Baang platform is voice-based, but its main features are fundamental to social media and shared across all major social media platforms. The content on Baang is generated by users in general, and this decentralization of content creation is an essential characteristic of social media platforms. Baang also has all of the standard mechanisms that allow users to engage directly with each other's content on social media: they can comment on, share, like and dislike each other's posts. Since all of the posts on Baang are also public and anonymous, the structure of Baang is directly analogous to the main page of reddit, one of the largest social media platforms with 52 million daily active users, and over 30 billion views every month (Curry, 2023). Furthermore, the users of Baang are in a demographic of young men with modest levels of education, which is of particular policy interest in developing countries, given their potential role in influencing a country's political and social stability (World Bank, 2006).

A simple framework generates three main hypotheses that structure the results in this paper. In the framework, a social welfare maximizing social media operator chooses a high or low level of moderation to maximize net information exposure (i.e. good exposure net of bad). A user then chooses their levels of exposure to good and bad information, which are determined by their total exposure to the platform as well as their perceptions of the relative quality of the sources of good and bad information.

The first prediction of our framework is that fully controlling access to information, as in both of the treatments, limits the overall usage of the platform if users have a distaste for moderation. Thus, we begin by documenting that the treatments reduce usage of the platform on both the extensive and intensive margins. On average, the treatments have 18% fewer daily users and are used for 41% fewer total daily minutes than the control. Thus, conditional on calling in, users spend 26% fewer minutes on the platform.

The second prediction is that if overall usage of the platform declines, then exposure to good information will also decline. This is confirmed by our first main result. We find that there is meaningfully *less* dissemination of official information in the treatments relative to the control condition. This result also holds for useful user-generated posts, whose content is aligned with the official information. On average, users in the treatments listen to 25% fewer minutes of official posts and 38% fewer minutes of aligned user-generated information than users in the control condition, although the absolute magnitude of the treatment effect on official information is much larger.

Next, we examine the *net* impact of high moderation on the dissemination of information. To do so, we compare exposure to all official and useful information net of misinformation under treatment as opposed to control. In the remove treatment, net information exposure *declines* by 21%. Thus, even though we remove all of the misinformation from the platform in that treatment, the decline in exposure to official information is meaningfully larger than the decline in exposure to misinformation. In the sunshine treatment, including rebuttals as a source of official information, the decline in net information exposure is 29%. These effects are largely driven by exposure to official information.

The third prediction of our framework explains the above result. It proposes that, given a distaste for moderation, high moderation will have a negative impact on net exposure if users' have a more favorable perception of the source of official information compared to misinformation. That is the case in this setting, where 95% of users indicate that they trust official Baang posts more than user-generated posts. According to the framework, these perceptions matter because they determine users' relative exposure to good and bad information, for a given level of total exposure. In this experiment, in the control, the average official post is listened to 653.2 more times and shared 62.5 more times than the average misinformation post, which is listened to 11.0 times and shared 0.0 times.

Finally, we examine the mechanisms that are driving the distaste for ex-ante moderation in this setting. Ex-ante moderation has two implications for how users experience the platform. The first implication is that it changes users' exposure to misinformation. One reason that users may have a distaste for moderation is if they have a preference for being exposed to misinformation. In that case, exposure to misinformation may increase engagement with the platform. Thus, we examine how users respond to being exposed to misinformation posts compared to how they respond to being exposed to matched user-generated posts that do not contain misinformation. In the control, users actually engage with the platform relatively more after being exposed to misinformation. In the sunshine treatment, however, there is no difference in users' engagement with the platform

after being exposed to misinformation, suggesting that exposure to rebuttals may mitigate this effect.

Another reason that users may have a distaste for ex-ante moderation is that it causes all content to be posted with a modest delay. This is the cost of a platform that is free of misinformation. In fact, all moderation leads to delays in identifying and addressing misinformation, even on major social media platforms.[4] The delays in this experiment are relatively modest. The difference between ex-ante and ex-post moderation is how that delay is resolved. Most major platforms rely on ex-post moderation, and thus that delay leaves misinformation on the platform for some period of time and allows it to be disseminated. A distaste for ex-ante delays in the posting of content will influence behavior under ex-ante moderation, and the effects should be concentrated after posting as users experience those delays. Using an event study approach, we demonstrate that for users who post, treatment effects are concentrated in the period after their initial post.

Our framework also highlights the settings in which the impact of ex-ante moderation on net information exposure would be positive, instead of negative as we observe here. In particular, users' relative exposure to official information as opposed to false information under low moderation matters. In this setting, relatively high levels of trust in the official information is likely instrumental in inducing users to seek out that information at high rates. This is in contrast to previous research on Twitter, for example, which finds that misinformation has a higher level of engagement relative to other types of information (Vosoughi, Roy and Aral, 2018). Thus, the framework indicates that even if users have a distaste for moderation, relatively high levels of engagement with misinformation can make ex-ante moderation optimal.

This is the first publicly available experiment that fully controls access to misinformation across an entire social media platform.[5] This design uniquely allows us to consider the implications of fully controlling for misinformation on the dissemination of all types of information. The few previous experiments on misinformation and social media have relied on controlled environments, and have tested the impact of exposing people to an individual piece of misinformation as well as various approaches to addressing that misinformation (Barrera et al., 2020; Henry, Zhuravskaya and Guriev, 2021; Pennycook et al., 2020). In addition, this study is related to concurrent studies on the moderation of toxic content. Jiménez Durán (2021) finds minimal effects of moderation on those being moderated directly, while Beknazar-Yuzbashev et al. (2022) finds evidence for a distaste for moderation effect as we do in this experiment. The experiment presented in this paper, however, is unique in examining the impact of moderation in the context of disseminating

[4] See Section 2 for further details.

[5] We know that social media companies experiment widely, but they do not always share the results of those experiments publicly.

official information, which has important policy implications. It also considers the implications of moderating misinformation specifically, as opposed to toxic content. Furthermore, although these results are broadly relevant, this experiment is the first to examine questions of misinformation or moderation in a development setting.

This experiment is also related to a broader literature in economics on how social media can expose people to information and other types of persuasive content. Thus far, it has largely examined the impact of social media on political attitudes and outcomes (Zhuravskaya, Petrova and Enikolopov, 2020; Fujiwara, Müller and Schwarz, 2021; Enikolopov, Makarin and Petrova, 2020). This literature largely relies on natural experiments, an exception is Levy (2021) who conducts an experiment to examine how the dissemination of news sources on social media affects attitudes. Another thread has measured the impact of reducing exposure to social media on exposure to information (Allcott et al., 2020; Mosquera et al., 2020). A few recent experiments have examined the potential to use social media to disseminate useful health information through third-party advertising or influencers (Breza et al., 2021; Alatas et al., 2021). In this setting, however, the source of useful health information is the platform itself.

2 Conceptual Framework

A simple two-period framework formalizes the hypotheses we test in this experiment. In the first period, a social welfare maximizing operator of a social media platform chooses either a high or low level of moderation m_k for $k \in h, l$ in order to maximize a user's positive net exposure to information $G = g - \theta b$, where g is the user's exposure to official, good information on the platform, and b is their exposure to user-generated misinformation. The operator could also assign a weight, θ, if they believe that the harm from bad information is greater than the benefit from good information, or vice versa.[6] In the second period, the user chooses their total exposure to the platform, $t(m_k)$, which is a function of the level of moderation. The user's exposure to both good $g(t(m_k), p_g)$ and bad $b(t(m_k), p_b)$ information is a function of t. It is also a function of p_j, the user's perceptions about the quality of a given type of information for $j \in g, b$, with relatively higher levels of perceived quality of an information source inducing relatively higher consumption.[7] Thus, net exposure is given by: $G_k = g_k(t(m_k), p_g) - \theta b_k(t(m_k), p_b)$. Using backward induction, the operator

[6]Note, we abstract away from the quantity of good or bad posts on the platform and rather focus on the time users spend listening to those posts. Individual posts on social media can have dramatically varying levels of reach and thus the quantity of posts is likely to be second order to the amount of exposure.

[7]We refer to perceptions, rather than beliefs here, since evaluating whether users update is not in the scope of this study. It is intuitive that users will spend more time consuming information from sources perceived to be of higher quality. For example, someone who trusts the *New York Times* and has little trust in Fox News is likely to much more time consuming news from the former source, while someone with the opposite perceptions is likely to spend their time in a way that is reversed.

will choose a level of moderation by comparing G_l and G_h.

A functional form simplification helps to more clearly illustrate the hypotheses generated by this framework. Specifically, let $g_k = p_g * (t(m_k))$ and $b_k = p_b * (t(m_k))$, where the perceptions regarding the sources of good and bad content are simply the probabilities of seeking out the two types of content. Then, $G_l > G_h$ when $(p_g - \theta p_b)t'(m_k) < 0$. That is, low moderation will be optimal when $t'(m_k)$, which is the preference for moderation, and $p_g - \theta p_b$, which is the relative exposure to good as opposed to bad information, have the different signs. We consider potential mechanisms that can determine the sign of $t'(m_k)$ in Section 6.[8]

The framework generates the three main hypotheses that we focus on testing in this paper. First, we can determine the sign of $t'(m_k)$ by measuring the impact of treatment on overall usage of the platform. Second, the direction of the change in g from control to treatment will be determined by the sign of $t'(m_k)$.[9] The third hypothesis is concerned with the conditions under which $G_l > G_h$. Specifically, if $t'(m_k) < 0$, then it must be the case that $p_g - \theta p_b > 0$.[10] That is, our third hypothesis is that, conditional on a distaste for moderation, low moderation will be optimal if the perception of, or trust in, the source of official information is greater than that of misinformation.

This framework also has broader relevance. In particular, it highlights in what contexts high, rather than low, levels of moderation are optimal. In particular, $p_g - \theta p_b$ and $t'(m_k)$ should have the same sign. Thus, even when users have a distaste for moderation, if there is more bad information than good information on the platform, the social welfare maximizer will choose high moderation.[11] In addition, it is straightforward to extend the framework to platforms where users are not anonymous. Although in this setting, official information comes only from the platform, it could also come

[8]We focus on the social planner case in this framework, since it is the first order question from a policy perspective. In addition, this platform is not for profit. Furthermore, the objective of the profit maximizing social media platform operator is to maximize engagement. In the context of this framework, that implies maximizing total exposure: $T_k = g_k(t(m_k), p_g) + \theta b_k(t(m_k), p_b)$. Then, it is straightforward to see that a profit maximizer will choose the moderation level based only on users' distaste for moderation.

[9]Note we do not have a general hypothesis here concerning b since it varies across treatments, and it is zero by construction in the remove treatment. Furthermore, our data clearly supports a linear relationship for $g_k = p_g * (t(m_k))$. Whether $b_k = p_b * (t(m_k))$ is more difficult to test given the specifics of our study design. The linearity assumption is reasonable, but a more general model could allow p_b to vary according to the level of moderation. Instead, for this experiment, we simply allow the remove treatment to be a special case in which $b_k = 0$.

[10]Considering how policymakers should set θ is beyond the scope of this paper, and thus we will set it equal to one in our analysis. Note that $p_g - p_b > 0$ is a necessary for $G_l > G_h$, when $t'(m_k) < 0$. In the remove treatment, however, a further restriction is required for $G_l > G_h$, specifically, $p_g t'(m_k) + p_b * t(m_l) < 0$. That is, the absolute magnitude of $p_g t'(m_k)$, which is the decline in good information from treatment to control, must be larger than $p_b * t(m_l)$, which is total reduction in misinformation from control to treatment in that case.

[11]Note that if users have a preference for moderation, then high moderation will be optimal when $p_g - \theta p_b > 0$. In addition, in the special case of the remove treatment, if there is a preference for moderation, then high moderation is always optimal.

from official government accounts or other trusted sources.[12] In that case, p_g and p_b would be the weighted averages of the perceptions of the sources of good and bad information.

The predictions regarding the impact of sunshine as opposed to remove are ambiguous. The sunshine treatment will expose users to more misinformation relative to the remove treatment. It can also, however, expose users to more official information in the form of rebuttals. Unlike the official posts, which users must seek out, the users may come across the rebuttals in the course of listening to standard feeds.

This framework focuses on user exposure to information, since that exposure represents important choices that users make with regards to their information-seeking behavior. Users' perceptions, however, could further lead them to weigh some sources of information more highly than others per unit of exposure. For example, higher levels of trust in official as opposed to user-generated information could lead users to not only increase their relative exposure to good information, but also give that information more weight in forming beliefs.[13]

3 Context

In the context of a public health crisis, disseminating health information widely and quickly is an important policy challenge in all types of countries. It is particularly relevant in the context of low-income countries such as Pakistan which are characterized by limited access to health information as well as health systems with limited capacity (Kremer and Glennerster, 2011; Dupas and Miguel, 2017). Holding other factors equal, the limited capacity of the health system increases the mortality risk from any outbreak. For example, Pakistan has 6.3 hospital beds per 10,000 people compared to the global average of 27.9 (WHO, 2021).

The study was implemented in the context of Baang, a non-profit, voice-based social media platform in Pakistan (Raza et al., 2018, 2022).[14] Voice-based social media platforms have relevance in many development contexts since they can be used by low-literate populations and those without an internet-connected phone or computer. Such platforms reach millions of users, especially in India, with Mobile Vaani being the most prominent example (Moitra et al., 2016). While these platforms primarily consist of user-generated content, like Baang, they also often aim to address particular information gaps, like the official posts on Baang did early in the COVID-19 pandemic.

[12] On such platforms, users regularly see and make decisions about what content to engage with based on its source (Levy, 2021).

[13] If prior beliefs about specific sources of misinformation are formulated outside the model, it may be more effective to directly rebut misinformation. That would make the sunshine treatment optimal *if* exposure to misinformation and official information is constant across the remove and sunshine treatments.

[14] Baang means rooster call in Urdu.

Past examples have focused on promoting citizen journalism (Marathe et al., 2015), agricultural information exchange among farmers (Patel et al., 2010), connecting employers and employees in rural settings (White et al., 2012), and allowing people in rural areas to ask questions of community health workers (Sherwani et al., 2007).

3.1 Baang platform

When users call into Baang, they are presented with a menu that gives them the option to: (1) listen to official Baang posts about COVID-19, (2) record their own posts, (3) listen to others' posts, or (4) listen to their own previously recorded posts.[15] After selecting option (3), users can then choose how they listen to others' posts, by: newest, today's most liked, or overall most liked. After each post plays, users are given the option to record an audio comment, listen to existing audio comments, forward (i.e. share), like, dislike or flag the post before moving on to the next post in the stream.[16] At any point while listening to posts, users can skip to the next post, and they frequently take advantage of this option. Option (1) is identical to (3) in that users are presented with a stream of posts and can comment on and engage with those posts, except (1) only includes the seven official Baang posts about COVID-19.[17] Finally, all of the users of Baang are anonymous, in so far as there are no public identifiers for each user, and all of the posts are public.

Thus, Baang has the fundamental characteristics of major social media platform, such as Facebook or Twitter (Boyd and Ellison, 2007). Users can generate and publicly share content, and then they can engage with other's content in a number of ways. Furthermore, posts on Baang does do receive substantial engagement; the average post receives 4.4 comments, 6.4 shares, and 7.2 likes.[18] Of the major social media platforms, Baang is most analogous to the main page of reddit, for some time called 'the front page of the internet' (Singer et al., 2014). reddit is a popular social media platform with 5 (2) billion monthly visits from across the globe (U.S.), and is the 7th (4th) most visited website in the world (U.S.), with a similar popularity to Instagram (Semrush, 2023). The format of Baang is also similar to browsing the "trending topics" page on Twitter, which includes posts on the most popular topics on Twitter at any given time.

Furthermore, much of the content on Baang is typical of other social networks. People often share and comment on the news, or tell personal stories. A number of users leverage the audio nature of

[15] Across Baang, options are selected by pressing numbers on users' phones.

[16] Users that choose to forward a post are asked to input the phone numbers of those they wish to forward to. Each inputted phone number is then sent an SMS message inviting them to call into Baang to listen to the post forwarded to them, indicating the forwarding user's phone number. If the user forwarded a post calls in following the SMS invitation, they are taken straight to the forwarded message before being sent to the main menu.

[17] See Section 4 for more details on the official Baang COVID-19 posts.

[18] See Section 4.2 for more on the usage of the platform.

the platform to recite or sing religious poetry, while others post self-described "radio shows" at the same time daily, to try to engage a regular audience. Of course, unlike a radio show or podcast, any user can set their own topic of discussion by posting. They can also respond to others' posts without mediation by engaging with them through likes, dislikes, or they can simply skip to the next post.

Prior to 2020, Baang had at times been a subsidized platform, in that the platform paid for the airtime of users while they were on the platform.[19] During a subsidized deployment in 2015, the platform reached more than 10,000 users over eight months through organic spread. These users actively engaged with the platform by calling in 293,657 times to participate through 35,677 posts and 155,352 comments. The posts were played 2.5 million times. In the months prior to the RCT, however, the platform was not subsidized. Thus, it had a smaller number of committed users, with 392 calling in on a typical day before treatment began.

4 Experiment Design

4.1 Timeline

An unsubsidized version of Baang was running prior to the COVID-19 pandemic. In April 2020, we made available official COVID-19 posts to all users on the platform. We then conducted the content moderation experiment for two months, from June 27th to August 26th, 2020. In addition, the day the randomized experiment began, we began to subsidize Baang to encourage the user base to grow. The platform was only completely free, however until July 25th. After that, due to funding limitations, users only had 30 free minutes a day to use the platform, with the potential to gain some additional minutes by forwarding the platform.[20] These adjustments to the cost of the using the platform were the same across all conditions, and thus were orthogonal to treatment.

4.2 Platform usage

During the randomized experiment, the platform generated meaningful engagement, from a total of 3698 users.[21] In total, users called into the platform 116,124 times. In addition to listening to content, the users recorded 13,315 posts and 69,768 comments. They further participated through 109,844 likes and 96,693 shares. Over this time period, the platform had 583 average daily users,

[19] As in most developing countries, in Pakistan, people typically pay for cell phone airtime by the minute, and purchase it in relatively small amounts at a time.

[20] Free minutes accrued across days. In addition, the option to gain additional minutes by forwarding the platform began on July 30th, and the number of minutes was increased on August 13th.

[21] Of those, 43% called in during the pre-experiment period that started in April, and the remainder called in for the first time during the experiment itself.

with the mean (median) user spending 23 (6) minutes on the platform per day.

User-generated COVID-19 content was a relatively small part of the total platform, which is perhaps not surprising given the demographics of the Baang user base. During the experiment, users generated 389 COVID-19 related posts and 532 COVID-19 related comments. This is approximately 1.1% of the total content on the platform (2.9% of posts and 0.7% of comments). Still, the total engagement with this content was substantive. Users spent 3886 minutes listening to user-generated COVID-19 posts, which generated 1380 comments, 294 shares, and 2034 likes.

4.3 Survey and user characteristics

Several months after the experiment, we conducted a phone survey on a sub-sample of 259 Baang users. This allowed us to learn users' demographics as well as their perceptions of different sources of information, including official Baang posts.[22]

The user base of Baang is largely younger males with modest education levels.[23] The average user is 30 years old, and around half (47%) have less than 10 years of education (Table 4). One-fifth have less than 8 years have education, and thus never reached upper secondary. Given that voice-based platforms are designed to be accessible to people without smartphones, a higher than expected percentage of users have a smartphone (91%). In addition, almost all of those (96%) regularly use WhatsApp, a common form of social media in this setting.[24] This suggests the potential broader relevance of voice-based platforms, since most users could spend their time on other higher profile social media platforms, but use Baang anyway.

4.4 Interventions

Before the experiment started, we added the official posts about COVID-19 to the platform. They were introduced with a clarification that they were official posts from Baang and were based on recommendations from local official sources such at the NIH, Pakistan. These seven posts stayed the same for the duration of the experiment and totalled approximately 6.5 minutes of content. The first sentence or two of the post contained its main message so that critical information in the post would reach users who did not listen to the entire post.

[22]The survey took place in April 2021 and included three samples: 94 randomly sampled users, 87 of the most active users, and 86 of the users most exposed to misinformation. Since these three groups of Baang users all have similar characteristics in practice, we focus on the randomly sampled users in the analysis discussed in the paper. See Section 5.2.1 for more on perceptions.

[23]That women mostly do not participate is perhaps unsurprising given the barriers to female participation in some aspects of public life in Pakistan (Schwab et al., 2016).

[24]Although Whatsapp is a messaging application, in many countries it is also used as social media as users join large groups where they do not know the other members.

Given the very limited number and length of the official posts on the platform, the engagement they generated was substantial. During the experiment, users spent 2717 minutes listening to those seven posts. They also engaged with the official posts through 162 comments, 978 shares, and 489 likes. Overall, 34% of users listened to official posts during the experiment. This number is constrained by the fact that another 10% of users had already sampled the posts before the experiment began. Overall, these statistics suggest that had more official posts been made available during the study, the findings reported in this paper could have even been more pronounced.

Instead of relying on local sources, the content in the rebuttals largely relied on content from international sources, such as the WHO, which had published rebuttals to COVID-19 related myths at that time.[25] Both the official posts and rebuttals were recorded by a single professional voice artist to further help users identify the official content as such.

4.4.1 Treatments

This experiment tested three approaches to addressing misinformation on the platform using two treatments and a control. The two treatments relied on *ex-ante* moderation, which means that all user-generated content on the platform was reviewed by a moderator before being made publicly available. Much of the misinformation on the platform could be identified by relying on pre-existing lists of myths created by international public health authorities.[26] In the *remove* treatment, we never posted the content identified as misinformation related to COVID-19. In the *sunshine* treatment, we posted all of the identified misinformation content, but we included a specific rebuttal with each piece of content.[27] These rebuttals played automatically immediately after the misinformation content, and were identified as official responses from the platform.[28]

These two treatments are compared against a control condition that relied on *ex-post* community-based moderation. This approach to moderation is similar to that of many social media platforms, and it was the standard on Baang before the study began. In the control, all user-created content was available immediately as it was posted, but users could tag messages as potential COVID-19 misinformation. These tagged posts were then sent to moderators to remove from the platform if found to be misinformation.[29]

[25] See Section SA1.1 for further details.

[26] Most examples of misinformation in this setting were largely unambiguous and included folk cures for COVID-19 as well as various conspiracies about it. See Section SA1.1 for additional details about the moderation and the rebuttals.

[27] A meta-analysis of lab experiments in psychology finds that specific rebuttals are more effective than simply denying misinformation in causing people to update their beliefs (Chan et al., 2017).

[28] Before the experiment began moderators reviewed all existing content on the platform and removed all COVID-19 misinformation to ensure that the remove treatment, in particular, was truly free of misinformation once the experiment started.

[29] During the experiment, these posts were already being reviewed, but they were only taken down in the control if

It it important to note that users in all three conditions were exposed to the same content in general. The only two exceptions were that users in the remove treatment were not exposed to COVID-19 misinformation posts, and users in the sunshine treatment were exposed to the official rebuttals. Otherwise, whenever a user posted to the platform, regardless of that user's own condition assignment (remove, sunshine, or control), that post was available immediately to everyone in the control condition. The same post would only become available to users in the two treatment conditions, however, once it was moderated. In addition, we did not announce to treatment users that the content they were exposed to had been ex-ante moderated. If some users became aware that other users did not receive the announcement, it might have induced them to shift across conditions, threatening the internal validity of the study.[30]

4.5 Random assignment

We designed our randomization to account for networks of users.[31] Specifically, treatment assignment depended on how a user reached the platform for the first time. *Original* users, who called in directly, were randomly assigned to one of the three conditions when they called into the platform for the first time during the study period. *Referral* users, who called in because they were forwarded content from the platform by another user, were assigned to the same condition as the user who forwarded them content. Regardless of how a user came to the platform initially, once a user was assigned to a condition, they remained in that condition every time they called into the platform thereafter. Users were identified by phone number.[32]

Randomizing referral users into the same treatment as their original user was intended to account for potential spillovers across conditions. Although the clusters only partially captured sharing networks, any spillovers across condition assignment that we were unable to fully capture with this randomization design would generally work against us finding effects.[33] Thus, we do not expect that any cross-treatment sharing is driving our results. Furthermore, our results are robust to accounting for spillovers in the analysis.[34]

We assigned a latent treatment status to each user as they called in starting in April. Of course,

identified by the community in order to ensure that typical moderation protocols were maintained. Users flagged 459 posts as misinformation during the study, but none of them were deemed misinformation by the moderators, suggesting the limitations of relying on community moderators in this setting.

[30]Furthermore, both potential mechanisms we propose to drive a distaste for moderation in Section 6 do not require users to be aware that they are being ex-ante moderated.

[31]We confirm the validity of the randomization in Section SA2.1.

[32]For more on this see Section SA1.1.4.

[33]For example, if control users forward the official COVID-19 posts to users who are in one of the treatment conditions, that would reduce the impact of being assigned to the treatments on exposure to official COVID-19 misinformation.

[34]Section SA2.2 examines the extent to which the randomization accounts for spillovers.

until the experiment began in late June, all users were effectively in the control. Thus, it is useful to differentiate two types of referral users. *Pre-treatment referrals* first used the platform before the experiment began, and thus the referral could not have been endogenous to treatment. *Post-treatment referrals* could conceivably have been selected into a given condition, since their condition was assigned after the study began. Thus, we consider our results on two samples. The sample of original users and pre-treatment users had treatment assigned exogeneously. The results for the full experiment sample is also an object of interest, however. If one condition is attracting more people or people who listen to more content, that is relevant to understanding the implications of a that condition.

There are 2077 original and pre-treatment referral users. This is just 56% of the full experimental sample, but they account for most of the platform usage. These users made 91% of the calls, recorded 94% of the posts and made 95% of the comments.[35]

Since the condition assignment of an original user determines the assignment of their referral users, this study relies on cluster-level random assignment. Each of the clusters in the study includes no more than one original user and their referral users if any. The average cluster includes just a few users, and thus there are 1408 clusters in the full sample and 1259 in the sample of original and pre-treatment referral users. As designed, the original users are almost exactly split across treatment conditions, with 367, 366, and 371 users assigned to the remove, sunshine and control conditions respectively.[36]

4.6 Outcome data

All of the main analysis in this study, and the outcomes in particular, rely on data that is automatically collected in the platform log files as users interact with the platform. The main outcomes in the experiment examine exposure and engagement at the user-level for three sources of information. We particularly focus on the impact of the treatments on the *official* information posts, since these posts were designed to provide high-quality information about COVID-19.

We also examine two sources of user-generated information: useful and misinformation posts. *Useful* posts contained information about COVID-19 that is aligned with the content in the of-

[35] They account for 484 average daily users out of a total of 583, with the mean (median) user spending 25 (9) minutes on the platform per day. Forty-six percent of these users listened to official posts during the experiment and an additional 18% before the experiment began.

[36] Due to natural sampling variation, in the full sample, there are 1153, 1258, and 1287 users in the remove, sunshine and control conditions respectively. In the sample of original and pre-treatment referral users, there are 681, 672, and 724 users in the remove, sunshine, and control conditions respectively. In addition, note that total number of original users is 1104, which does not equal the number of clusters (1408) in the study. This is because in some cases an original user called in before the experiment began, referred the platform to someone else, and then never called in during the experiment itself. Thus, there are 304 clusters that do not include an original user.

ficial posts. In some cases, that included personal experiences with COVID-19 that emphasize that it is real. This type of content is aligned with one of the goals of the official content, which was to confirm that COVID-19 was not a hoax. *Misinformation* posts contained false information about COVID-19. User-generated COVID-19 posts were twice as likely to be useful (21%) as opposed to misinformation (8%). Useful information was identified and categorized after the experiment, while misinformation was identified during the experiment through moderation. This categorization was double-checked after the experiment.[37] Most user-generated posts about COVID-19 (71%), however, were neither useful not misinformation and thus were categorized as *neutral*.

For each of the three types of information, we conduct our analysis for three exposure and engagement measures. The focus of our analysis is on exposure to information, which is measured through minutes spent listening to a given type of post. Since users have a great deal of control on how they spend their time on the platform, this is the key measure of information-seeking behavior that is of interest here.

In addition, we consider two measures of engagement. Increased engagement can induce additional exposure of other users directly, through sharing, or indirectly through increasing the popularity of posts. It can also potentially characterize the intensity of users' exposure. We separately examine one measure of engagement, the number of shares, since it is the primary outcome of interest for researchers focusing on the determinants of the spread of misinformation. We also examine a standardized index of the other measures of engagement: comments, likes, and dislikes.

4.7 Estimation

The main results are intention-to-treat (ITT) estimates of the impact of the treatment at the user-level for the full study period. We initially present our results graphically, which exploits the time series nature of our data. We focus on the user-level analysis for the main results, however, because it eliminates considerations about attrition or selection over the course of the study. Everyone who is in the experiment, regardless of condition assignment, appears in the data once. Thus, when official information exposure is our outcome variable, for example, it is includes the total amount of exposure across all days that the user called in during the RCT.

Thus, the main estimating equation is given by:

$$Y_i = \beta_1 Remove_i + \beta_2 Sunshine_i + \epsilon_c,$$

[37] For more on content categorization, see Section SA1.1.

where $Remove_i$ is an indicator for having been assigned to the remove treatment and $Sunshine_i$ is an indicator for having been assigned to the sunshine treatment. Our other main estimates consider the effect of being assigned to either the sunshine treatment or remove treatment using the indicator $Treated_i$. In both cases, we cluster the standard errors at the level of an original user and their referral users. The mechanism results rely on a non-parametric event study approach, which is discussed in that section.

As noted above, we present our main results for both the sample of original and pre-treatment referral users and the full experimental sample. For other results, however, we focus on the former sample.

5 Results

The first hypothesis of our framework is that if users have a distaste for moderation, then usage will decline under the higher level of moderation in the two treatments. Thus, we document that both treatments reduce the overall usage of the platform (Figure 1 Panel A). We consider two extensive margin measures of usage: total number of users per day and total minutes per day. On average, the treatments attract 43.1 (19%) fewer daily users relative to the control. Those users spend a total of 2120 (42%) fewer minutes on the platform. The effect of treatment on the intensive margin measure of usage, minutes per user per day, is also substantive. Users assigned to the treatments spend an average of 5.6 (26%) fewer daily minutes on the platform. Thus, not only do fewer users in the treatments call in on any given day than in the control, but conditional on calling in, those users spend less time on the platform. We confirm the statistical significance of these results in Section SA2.1.

We also present an initial graphical representation of the second hypothesis, that a decline in overall usage of the platform will induce a decline in exposure to official information (Figure 1 Panel B). On the extensive margins, fewer users in the treatments are exposed to official posts on the average day and they collectively spend fewer minutes listening. Notably, although the probability of calling in a given day depends on treatment status, conditional on calling, users have the same likelihood (16%) of listening to official posts across conditions.[38] Thus, the main treatment effects outlined in the following subsection are likely to be driven by the extensive margin.

These figures illustrate that the impacts of being assigned to treatment are concentrated in the first half of the experiment. There are two reasons for that, both of which suggest that our results are a lower bound, however, on the potential impact of treatment. First, the official posts remained

[38]This finding supports the assumption of linearity in the framework, namely, that $g_k = p_g * t(m_K)$.

the same throughout the study, so once the users sample them, there would likely be diminishing marginal returns to repeat listens. Second, once access to the platform became limited, as indicated by the red line on the figures, the cost of spending time listening to all types of posts increased. The large resulting decline in overall usage makes detecting the impact of treatment more difficult.

5.1 Main results

Our main results are motivated by the second hypothesis of the framework, that a decline in overall usage of the platform will induce a decline in exposure to official information. In addition to examining exposure to official information, we also measure the impact of treatment on our measures of engagement, since they are potentially important in both influencing and characterizing exposure. The numbers reported below are for the sample of original and pre-treatment referral users, but the results are qualitatively similar for the full experimental sample, as is evident in the referenced tables.

Table 1 provides estimates of the impact of the two treatments on exposure to and engagement with official information posts. Since the results are similar for each of the two treatments, our focus is on their average effect. Users assigned to the treatments listen to 0.33 (25%) fewer minutes of the official posts relative to users assigned the control, a result which is significant at the 5% level.[39] In the control condition, users listen to an average of 1.33 minutes of the official posts, which is a somewhat more that one such post. The results on the engagement measures are the expected sign, but are marginally insignificant for the main specifications. This is unsurprising given that engagement is a subset of exposure, and thus it is more difficult to detect effects on these measures. They are significant, however, for some individual treatment effects. Thus, these findings are suggestive, but not definitive with regards to the impact of treatment on engagement.

Next, we examine the impact of being assigned to either of the two treatments on exposure to useful user-generated information.[40] The treatment effects on the useful posts are smaller in absolute magnitude but larger in relative magnitude that the effects on official posts (Table 2). Users assigned to either of the two treatments spend 0.14 (38%) less minutes listening to useful posts than in the control, which is significant at the 5% level. In the control, the exposure to useful

[39]One question that arises is whether users will seek the same knowledge elsewhere. Although answering that question is beyond the scope of this design, other recent work finds that reduced access to social media does not lead to seeking information from high quality sources and thus it reduces knowledge outcomes (Allcott et al., 2020; Mosquera et al., 2020). This suggests that, as hypothesized, users who get their news from social media are otherwise hard to reach.

[40]Note that useful information is aligned with the good information in our framework, and as discussed in that section it is possible to have multiple sources of good or bad information.

user-generated content in the control is 0.35 minutes, however, which is lower than for official posts. This highlights the reach of official content on the platform, which is further examined in Section 6 below. As is the case for the results on official posts, the average treatment effects on the engagement measures are at most marginally significant.

Finally, we examine the impact of treatment on exposure to misinformation (Table 3). We conduct this analysis separately by each treatment, since the two treatments had different objectives with regards to addressing misinformation. In the control condition, users are exposed to 0.126 minutes of misinformation on average. As intended, users in the remove treatment are exposed to effectively zero misinformation. Thus, the treatment effect on being assigned to the remove condition is negative 0.122, an 97% decrease relative to the control.[41] In the sunshine treatment, however, we do not see statistically significant differences in exposure to misinformation relative to the control.

In Section SA2, we document that these results are robust to accounting for outliers and spillovers.

5.2 Net information exposure

Motivated by our third hypothesis, we now test whether high moderation has a negative impact on net information exposure, or good minus bad information exposure. For the purposes of this analysis, we weight exposure to good and bad information equally, but we recognize policymakers may choose to weight differently depending on the context.

We begin examining this hypothesis by measuring the average effect of being assigned to one of the treatments on net exposure. Comparing official information to user-generated misinformation, we find that treatment decreases net exposure relative to the control by 25%. This approach to assessing net exposure is directly aligned with our model, which assumes that official and false information comes from different sources. We also measure net exposure including useful posts in the accounting of good information. Using that measure, we find that treatment decreases net exposure relative to the control by 29%. These findings capture that the absolute magnitude of the treatment effect is larger for good information than for bad information. Furthermore, they confirm that it is largely official information that matters in this setting.

Next, we examine average treatment effects on net exposure separately for the remove and sunshine treatments. In the remove treatment, users are not exposed to any misinformation, which may increase net exposure. In the sunshine treatment, however, users are exposed to the rebuttals, an additional source of good information not available to remove users. When we include only

[41] This exposure to misinformation is not exactly zero since a re-examination of all COVID-19 posts on the platform after the experiment was complete identified a small number of misinformation posts that were not identified initially.

official sources of good information in our measure (including rebuttals), we find that the remove treatment decreased net exposure relative to the control by 15% where the sunshine treatment does so by 25%. This pattern persists, with declines in net exposure of 21% and 29% respectively, if we include useful posts in our measure of good information. It is important to note that net exposure declines in the remove treatment even though we have removed of the misinformation from the platform. Another consideration in weighing tradeoff across sunshine and remove in a given context is whether the rebuttals are an opportunity to refute misinformation that is circulating outside the platform. We do find evidence that the misinformation on the Baang platform was disseminated through other sources in Pakistan (see Section SA3).

5.2.1 Conditions for negative net exposure

Given we find negative net information exposure from high moderation, we now examine the conditions of the third hypothesis. This hypothesis states that given a distaste for moderation, net exposure will be negative under high moderation if users perceive the sources of good information to be of higher quality than those of the false information. In our survey, we measure perceptions by asking users about their trust in various information sources. A random sample of users almost universally (95%) trust the official Baang posts over users' posts on COVID-19. This is also reflected in their responses to a separate set of questions, in which they are asked to rank their trust in COVID-19 information from different sources on a five-point scale. Their trust in official posts was 3.1 while their trust in users' COVID-19 Baangs was 2.2, a statistically significant difference (p-value 0.000). This range of 1.1 points on the scale also covers a large percentage of total observed range of trust levels in different sources of information. The least trusted source of information is users of other types of social media aside from Baang (1.8), while the most trusted sources are government announcements (3.8) and doctors (3.8).[42]

Perceptions of different information sources matter, according to the framework, since if users have greater levels of trust in official posts relative to user-generated misinformation posts, then they may be more likely to seek them out and engage with them. This would then lead to net negative information exposure under moderation. So, to complement the survey finding on trust, we directly examine user exposure to and engagement with different types of COVID-19 posts in the control group. Note that this analysis focuses on the intensity of exposure or engagement per post, and thus abstracts away from the quantity of different types of posts on the platform. Compared to misinformation posts, which are listened to 11 times on average, official posts are

[42]The Baang posts were aligned with government announcements, but it is perhaps unsurprising that they are less trusted since they are not a direct source. Trust in the official posts were in a similar range with trust in local imams (3.2).

listened to an 653 additional times.[43] The official posts are also shared 62.5 additional times and their engagement index is approximately 13σ greater than misinformation posts. The engagement index is normalized to zero for misinformation posts. Exposure to and engagement with useful information posts is modestly but significantly greater than that of misinformation posts, with users listening to such posts an additional 3.2 times more than the misinformation posts. This pattern is in contrast to studies in other settings where misinformation received more engagement than other types of posts (Vosoughi, Roy and Aral, 2018). [44]

6 Mechanisms

As outlined in the conceptual framework, whether users have preference for or distaste for moderation is a key determinant of the optimal approach to moderation from the perspective of a social planner. Using the platform more under high moderation indicates a preference for moderation, while using it less indicates a distaste for moderation. To better understand users' preferences over moderation, we consider two implications of high or ex-ante moderation in this setting.[45]

First, ex-ante moderation has implications for whether and how users are exposed to misinformation in this experiment. In the remove treatment, users are not exposed to misinformation, and in the sunshine treatment that exposure is mitigated by rebuttals. On the one hand, users may prefer moderation since they may not enjoy being exposed to misinformation. If that is the case, we may expect users to spend less time on the platform after being exposed to misinformation. On the other hand, users may have a distaste for moderation, in which case they may use the platform more after being exposed to misinformation since they enjoy such content. There have been recent high profile examples of both of these preferences.[46] Alternatively, exposure to misinformation

[43] See Table SA1. A listen is defined as ever beginning to listen to a post. There are structural reasons why listens might be higher for user-generated posts. Specifically, they play automatically in a user's feed, while a user would have actively seek out official posts. In addition, official posts are also a small percentage of the total platform, with just seven total posts. At the same time, however, the official posts are directly accessible throughout the study.

[44] One possible explanation for the low levels of engagement with misinformation this setting is that the official posts have an inoculation effect that increases scepticism towards misinformation, which could lead users to skip such posts. Roozenbeek et al. (2022) finds that exposure to information on rhetorical techniques has an inoculation effect against misinformation.

[45] Note although we will find evidence for distaste for moderation according to revealed preference through both of mechanisms proposed in this section, in surveys, Baang users do overwhelmingly indicate a preference for moderation. Specifically, 99% prefer that Baangs are moderated and, in a separate question, 100% prefer that they are moderated by the Baang team. The survey, however, does not indicate the details of moderation, thus users are likely considering ex-post moderation as opposed to no moderation. This is particularly likely given that WhatsApp is a common alternative form of social media in this setting and it doesn't have any moderation.

[46] On the one hand, when Twitter management explicitly stated they would be doing less moderation of misinformation the platform in late 2022, many users left the platform (Sweney, 2023). On the other hand, social networks such as Parler and Truth Social have made a stated lack of moderation their selling point, and have become a haven for prominent figures that had posts removed from Twitter due to misinformation (Lima, 2021).

may induce an emotional reaction (Brady et al., 2017; Lewandowsky and Van Der Linden, 2021; Vosoughi, Roy and Aral, 2018). This could lead to increased usage of the platform after exposure to misinformation if social media use is addictive or correlated with emotional disregulation (Sun and Zhang, 2021; Liu and Ma, 2019).

Second, ex-ante moderation is required in order for a platform free of misinformation, and this study highlights the cost of that approach for users. Specifically, ex-ante moderation requires that all user-generated content is posted to the platform with some delay in order to allow time for moderation. In this experiment, the average delay was a relatively modest 67 minutes, for the 99% of content that is not about COVID-19.[47] It is not surprising that even modest delays may affect usage. Social media usage patterns are consistent with understanding it as a reward system based on likes or engagement with one's own content (Lindström et al., 2021). Furthermore, social media use is correlated with delay discounting, and thus people who prefer instant rewards are more likely to be users (van Endert and Mohr, 2020).[48]

These delays are relevant to moderation in other social media platforms. In general, major platforms rely on ex-post moderation that only requires reviewing a fraction of the content on platform. In contrast, ex-ante moderation requires moderating all content. Despite this more limited approach, ex-post moderation is still significantly delayed on such platforms.[49] This is perhaps due to the costs of moderation, which include thousands of human moderators.[50] Under ex-post moderation, however, the implication of those delays is that people are exposed to misinformation.

Furthermore, moderating misinformation has unique challenges, and platforms have until recently largely focused on moderating more traditionally regulated toxic speech. New types of misinformation are always arising, and identifying misinformation can require significant additional time, since it typically requires expertise beyond that of standard moderators.[51] AI is likely to have a

[47] Covid-19-related posts not deemed misinformation had, on average, 270 minute delays to posting. Covid-19-related posts deemed misinformation had, on average, 554 minute delays. These increased delays were caused by our moderation process requiring a supervisor and in many cases a PI to sign off on such decisions.

[48] In practice, all posts go up the control immediately, and thus are equally likely to receive engagement. Users in the treatments, however, will not necessarily be aware of that engagement. In particular, we confirm that 44% users of who experience a delay check the main feed between the time they post and the time the post goes onto the platform. These users could expect to find their own post and be disappointed.

[49] There has been little attempt to comprehensively quantify overall moderation delays on major platforms, but Goldstein et al. (2023) find that average time to post removal on Facebook is approximately 21 hours in late 2020.

[50] Although figures are not reported publicly, according to some reports, Facebook relies on 10,000 to 15,000 human moderators (Barrett, 2020).

[51] Goldstein et al. (2023) finds content removal was significantly delayed after the U.S. capital riot on January 6th, 2021, as Facebook changed its policies to address new types of misinformation content. More generally, Facebook relies on an independent council to make determinations about what types of posted content is misinformation (Oversight Board, 2023).

growing role in moderation, but is also expected to increase the dissemination of misinformation, as well as the challenges in identifying it.[52] Furthermore the challenges of identifying misinformation and relying on AI moderation are even greater in development settings and in languages other than English.[53] Determining what is misinformation will take time and human judgement for the forseeable future.

We take an event study approach to test these potential mechanisms. This approach is appropriate here given the timing of the events are variable, and they can take place throughout the experiment. We also rely on a local polynomial regressions given the nonlinear nature of our data. In this analysis, we rely on a conditional parallel trends assumption for identification.[54]

6.1 Exposure to misinformation mechanism

In order to understand users' preferences for misinformation exposure, we examine user behavior after being exposed to misinformation. To do so, we take an event study approach in which the treatment event is exposure to a user-generated *misinformation* post for the first time and the counterfactual or control event is exposure to a comparable *useful or neutral* post. Given there are more non-misinformation posts than misinformation posts related COVID-19 during our study, we selected a matched subsample of non-misinformation posts to serve as counterfactuals identified through a propensity score approach.[55] The matching exercise simply selects the control group, however. For identification, in this event study framework, we confirm that usage before first user-generated COVID-19 post exposure follows the same trend. We conduct this analysis separately for the two conditions in which users are exposed to misinformation, the control and the sunshine treatment, since the rebuttals may impact user behavior.

We find that, in the control, users spend more time on the platform after exposure to misinformation posts relative to similar non-misinformation posts (Figure 2). This indicates a distaste for moderation on the part of users. Notably, we do not find that misinformation has the same effect on usage in the sunshine treatment. This suggests that rebuttals may have a mitigating effect on post-

[52] For a summary of the challenge of moderating misinformation on social media platforms and the role of AI in this process, see Gallo and Cho (2021). For a specific overview of the challenges of scaling content moderation through AI, see Gillespie (2020).

[53] According to documents released by a whistle-blower, the accuracy of a Facebook algorithm in detecting hate speech in the Afghan context was 0.2%. Furthermore, in Arabic, an algorithm falsely flagged innocuous content 77% of the time (Scott, 2021).

[54] We plot standard error bands to allow for visual examination of parallel trends.

[55] We matched on two post characteristics: date and number of total listens. This exercise compares subsequent engagement for users who came across a COVID-19 related user-generated post for the first time while listening to their feed, but for some users that post contains misinformation and other users it contained useful or neutral information. Given there was not a user-specific algorithm ordering feeds in our context, rather they come across a misinformation post or a useful/neutral post first is likely to be effectively random.

misinformation usage.[56] We also compare difference in the impact of misinformation exposure across the sunshine and control arms and find it is statistically significant (Figure SA3).

6.2 Delay mechanism

We hypothesize that the delay mechanism is most likely to be observable for users who post, since they have directly experienced their own content being delayed as it is posted to the platform.[57] Thus, in order to test whether this is a mechanism for the observed distaste for moderation, we examine whether treatment effects are concentrated during the time period after users post for the first time. Specifically, using an non-parametric event study, we examine both overall engagement with the platform and specific engagement with official posts as outcomes of interest. This analysis is limited to the subsample of 722 users who ever post from the sample of original and pre-treatment referral users, which allows us to examine pre-trends.[58] We verify that users in both the treatment and control have similar engagement with Baang until their first post.[59]

In Figure 3 Panel A, we find that, after a user's first post, treatment users use the platform significantly less than control users. Usage increases in both the treatment and control immediately after posting, which is consistent with evidence from other platforms (Grinberg et al., 2016). For users who have been assigned to the treatment, however, their usage declines more quickly that for those in the control. Although the confidence intervals are wide immediately after posting, the treatment effect persists from four days after posting to the end of our event study. In Figure 3 Panel B, we find a similar pattern for exposure to official posts specifically, further confirming our result. Finally, we do not find evidence that these results are particularly sensitive to the length of the delay, which is not surprising given users are likely to notice even short delays in their posts reaching the platform.[60]

[56] If users like misinformation, they may dislike the rebuttals and be driven off the platform by them. Alternatively, if the misinformation induces an emotional response, the rebuttals may mitigate that response.

[57] This is in contrast to users who spend time on the platform simply listening to content, and who are likely to be less aware of the fact that the content they are listening to is being posted with a modest delay.

[58] Focusing on this subsample allows us to examine whether pre-trends are parallel, which is a test of the identifying assumption for this approach. The pre-period is not defined for those who never post, since the timing of posting for the first time is variable and defined at the individual level for those who post. Thus it is less straightforward to test an identifying assumption for the subsample who never post, and we exclude these users from this analysis. We expect that users who post are a selected subsample, and thus, this analysis does not allow us to determine whether those who never post are affected by being assigned to treatment or not.

[59] This is not surprising since users are randomized into treatment groups initially, so if they are largely not treated before they post for the first time, then the determinants of the outcome would be similar in the pre-period for people who post.

[60] See Section SA2.5 for analysis of this question.

7 Conclusion

We conduct the first randomized controlled trial with publicly available results in which the two treatments aim to fully control (mis)information across an entire social media platform. In this case, we focus on information specific to COVID-19. We combine this experiment with an intervention that provides access to official posts that contain high-quality information regardless of treatment assignment. We document that a substantial percentage of users seek out these high-quality official posts on the platform (44%). We find that fully controlling access to misinformation through high or ex-ante moderation reduces usage of the platform. This leads to meaningfully less exposure to the official posts. We consider the impact of the treatments on net exposure, and find that they substantially reduce net exposure in to information (good minus bad).

The remove treatment in particular, is designed to understand the implications of a social media platform that is free of misinformation. Specifically, in the remove treatment, there is no COVID-19 misinformation on the platform. The decline in exposure to official information is greater than the decline in misinformation under treatment, however, even though we have removed all of the misinformation from the platform.

A conceptual framework helps to contextualize these results. It identifies that users' preference for moderation and their relative trust levels in good as opposed to bad information will determine the optimality of low or high moderation. In this setting, almost all users trust the official posts more that user-generated COVID-19 posts. This may explain why users are much more likely to listen to and engage with official posts compared to user-generated posts on the same topic. Furthermore, higher levels of trust in official posts may lead users to give the official information more weight than that from the misinformation posts. In contrast to this setting, in settings with lower levels of trust in official information and more dissemination of misinformation, ex-ante moderation is likely to be optimal.

In this experiment, users exhibit a distaste for moderation. Thus, we examine two potential mechanisms for that distaste, and we find evidence for both. First, users engage with the platform more after being exposed to misinformation. So, it is not surprising that they use the platform less after that type of information is removed. Second, this experiment highlights that a cost of ex-ante moderation is modest delays in user-generated content being made available on the platform, and that users dislike delays. These delays are uniquely relevant to moderating misinformation in particular. Given that there are significant delays in the ex-post moderation of content on major platforms in the status quo, these delays are also relevant to those settings.

Large social media platforms have not embraced fully eliminating misinformation through pre-

moderation, likely because of the challenges outlined in this paper. In contrast to this setting, however, large platforms often grapple with a context in which users engage with and share misinformation more than other types of information. To minimize the potential harm from those limitations, social media companies can take a two-pronged approach. First, platforms can step up efforts to actively disseminate high-quality information from trusted sources, and work to increase trust in reliable information sources. Second, they can continue to limit the spread of misinformation.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining causal findings without bias: Detecting and assessing direct effects." *American Political Science Review*, 110(3): 512–529.

Alatas, Vivi, Arun Chandrasekhar, Markus Mobius, Benjamin Olken, and Cindy Paladines. 2021. "Designing Effective Celebrity Messaging: Results from a Nationwide Twitter Experiment promoting Vaccination in Indonesia." Stanford working paper.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social media and fake news in the 2016 election." *Journal of Economic Perspectives*, 31(2): 211–36.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The welfare effects of social media." *American Economic Review*, 110(3): 629–76.

Barrera, Oscar, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya. 2020. "Facts, alternative facts, and fact checking in times of post-truth politics." *Journal of Public Economics*, 182: 104123.

Barrett, Paul M. 2020. "Who moderates the social media giants." *Center for Business*.

Beknazar-Yuzbashev, George, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski. 2022. "Toxic content and user engagement on social media: Evidence from a field experiment." *Available at SSRN*.

Boyd, Danah M, and Nicole B Ellison. 2007. "Social network sites: Definition, history, and scholarship." *Journal of computer-mediated Communication*, 13(1): 210–230.

Brady, William J, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. "Emotion shapes the diffusion of moralized content in social networks." *Proceedings of the National Academy of Sciences*, 114(28): 7313–7318.

Breza, Emily, Fatima Cody Stanford, Marcella Alsan, Burak Alsan, Abhijit Banerjee, Arun G Chandrasekhar, Sarah Eichmeyer, Traci Glushko, Paul Goldsmith-Pinkham, Kelly Holland, et al. 2021. "Effects of a large-scale social media advertising campaign on holiday travel and COVID-19 infections: a cluster randomized controlled trial." *Nature Medicine*, 27(9): 1622–1628.

CDC. 2018. *Crisis Emergency Risk Communication Manual.*

Chan, Man-pui Sally, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. "Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation." *Psychological Science*, 28(11): 1531–1546.

Curry, David. 2023. "Reddit Revenue and Usage Statistics (2023)." *Business of Apps*. January 9, 2023.

Dupas, Pascaline, and Edward Miguel. 2017. "Impacts and determinants of health levels in low-income countries." In *Handbook of Economic Field Experiments*. Vol. 2, 3–93.

Dwoskin, Elizabeth. 2020. "Facebook launches one-stop shop portal for coronavirus information." *Washington Post*. March, 18th, 2020.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova. 2020. "Social media and protest participation: Evidence from Russia." *Econometrica*, 88(4): 1479–1514.

Fujiwara, Thomas, Karsten Müller, and Carlo Schwarz. 2021. "The effect of social media on elections: Evidence from the United States." *National Bureau of Economic Research Working Paper 28849*.

Fung, Brian, and Devin Cole. 2023. "Biden administration defends communications with social

media companies in high-stakes court fight." *CNN*. August, 10th, 2023.

Gallo, Jason A, and Clare Y Cho. 2021. "Social Media: Misinformation and content moderation issues for Congress." *Congressional Research Service Report*, 46662.

Gillespie, Tarleton. 2020. "Content moderation, AI, and the question of scale." *Big Data & Society*, 7(2): 2053951720943234.

Goldstein, Ian, Laura Edelson, Damon McCoy, and Tobias Lauinger. 2023. "Understanding the (in) effectiveness of content moderation: A case study of facebook in the context of the us capitol riot." *arXiv preprint arXiv:2301.02737*.

Grinberg, Nir, P Alex Dow, Lada A Adamic, and Mor Naaman. 2016. "Changes in engagement before and after posting to facebook." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 564–574.

Henry, Emeric, Ekaterina Zhuravskaya, and Sergei Guriev. 2021. "Checking and Sharing Alt-Facts." *American Economic Journal: Policy*, forthcoming.

Hsu, Tiffany. 2023. "Falsehoods Follow Close Behind This Summer's Natural Disasters." *New York Times*. August, 30th, 2023.

Jiménez Durán, Rafael. 2021. "The Economics of Content moderation: Theory and experimental evidence from hate speech on Twitter." *Available at SSRN 4044098*.

Kremer, Michael, and Rachel Glennerster. 2011. "Improving health in developing countries: evidence from randomized evaluations." In *Handbook of Health Economics*. Vol. 2, 201–315.

Levy, Ro'ee. 2021. "Social media, news consumption, and polarization: Evidence from a field experiment." *American Economic Review*, 111(3): 831–70.

Lewandowsky, Stephan, and Sander Van Der Linden. 2021. "Countering misinformation and fake news through inoculation and prebunking." *European Review of Social Psychology*, 32(2): 348–384.

Lima, Cristiano. 2021. "Gettr, Parler, Gab find a fanbase with Brazil's far-right." *Washington Post*. November 9, 2021.

Lindström, Björn, Martin Bellander, David T Schultner, Allen Chang, Philippe N Tobler, and David M Amodio. 2021. "A computational reward learning account of social media engagement." *Nature Communications*, 12(1): 1311.

Lin, Xialing, Patric R. Spence, Timothy L. Sellnow, and Kenneth A. Lachlan. 2016. "Crisis communication, learning and responding: Best practices in social media." *Computers in Human Behavior*, 65: 601–605.

Liu, Chang, and Jian-Ling Ma. 2019. "Adult attachment style, emotion regulation, and social networking sites addiction." *Frontiers in Psychology*, 10: 2352.

Marathe, Megh, Jacki O'Neill, Paromita Pain, and William Thies. 2015. "Revisiting CGNet Swara and its impact in rural India." *Proceedings of the Seventh International Conference on Information and Communication Technologies and Development*, 1–10.

Moitra, Aparna, Vishnupriya Das, Gram Vaani, Archna Kumar, and Aaditeshwar Seth. 2016. "Design lessons from creating a mobile-based community media platform in Rural India." *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, 1–11.

Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie. 2020. "The economic effects of Facebook." *Experimental Economics*, 23(2): 575–602.

Oversight Board. 2023. "Oversight Board." http://www.oversightboard.com, accessed August

2023.

Patel, Neil, Deepti Chittamuru, Anupam Jain, Paresh Dave, and Tapan S Parikh. 2010. "Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 733–742.

Pennycook, Gordon, Adam Bear, Evan T Collins, and David G Rand. 2020. "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings." *Management Science*, 66(11): 4944–4957.

Raza, Agha Ali, Bilal Saleem, Shan Randhawa, Zain Tariq, Awais Athar, Umar Saif, and Roni Rosenfeld. 2018. "Baang: A viral speech-based social platform for under-connected populations." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Raza, Agha Ali, Mustafa Naseem, Namoos Hayat Qasmi, Shan Randhawa, Fizzah Malik, Behzad Taimur, Sacha St-Onge Ahmad, Sarojini Hirshleifer, Arman Rezaee, and Aditya Vashistha. 2022. "Fostering Engagement of Underserved Communities with Credible Health Information on Social Media." *Proceedings of the ACM Web Conference 2022*, 3718–3727.

Roozenbeek, Jon, Sander van der Linden, Beth Goldberg, Steve Rathje, and Stephan Lewandowsky. 2022. "Psychological inoculation improves resilience against misinformation on social media." *Science Advances*, 8(34): eabo6254.

Schwab, K, R Samans, S Zahidi, et al. 2016. "The Global Gender Gap Report 2016: World Economic Forum."

Scott, Mark. 2021. "Facebook did little to moderate posts in the world's most violent countries." *Politico*. October, 25th, 2021.

Semrush. 2023. "Top Websites." http://www.semrush.com, accessed August 2023.

Sherwani, Jahanzeb, Nosheen Ali, Sarwat Mirza, Anjum Fatma, Yousuf Memon, Mehtab Karim, Rahul Tongia, and Roni Rosenfeld. 2007. "Healthline: Speech-based access to health information by low-literate users." 1–9, IEEE.

Singer, Philipp, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel, and Markus Strohmaier. 2014. "Evolution of reddit: from the front page of the internet to a self-referential community?" *Proceedings of the 23rd international conference on world wide web*, 517–522.

Stiglitz, Joseph E. 2000. "The contributions of the economics of information to twentieth century economics." *Quarterly Journal of Economics*, 115(4): 1441–1478.

Sun, Yalin, and Yan Zhang. 2021. "A review of theories and models applied in studies of social media addiction and implications for future research." *Addictive Behaviors*, 114: 106699.

Sweney, Mark. 2023. "Twitter 'to lose 32m users in two years after Elon Musk takeover'." *The Guardian*. December 13, 2022.

Tursunbayeva, Aizhan, Massimo Franco, and Claudia Pagliari. 2017. "Use of social media for e-Government in the public health sector: A systematic review of published studies." *Government Information Quarterly*, 34(2): 270–282.

Twitter. 2020. "COVID-19 tab in Explore." https://blog.twitter.com/en_us/topics/company/2020/covid-19#explore.

van Endert, Tim Schulz, and Peter NC Mohr. 2020. "Likes and impulsivity: Investigating the relationship between actual smartphone use and delay discounting." *PloS One*, 15(11): e0241383.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The spread of true and false news online." *Science*, 359(6380): 1146–1151.

Wang, Yuxi, Martin McKee, Aleksandra Torbica, and David Stuckler. 2019. "Systematic lit-

erature review on the spread of health-related misinformation on social media." *Social Science & Medicine*, 240: 112552.

White, Jerome, Mayuri Duggirala, Krishna Kummamuru, and Saurabh Srivastava. 2012. "Designing a voice-based employment exchange for rural India." 367–373.

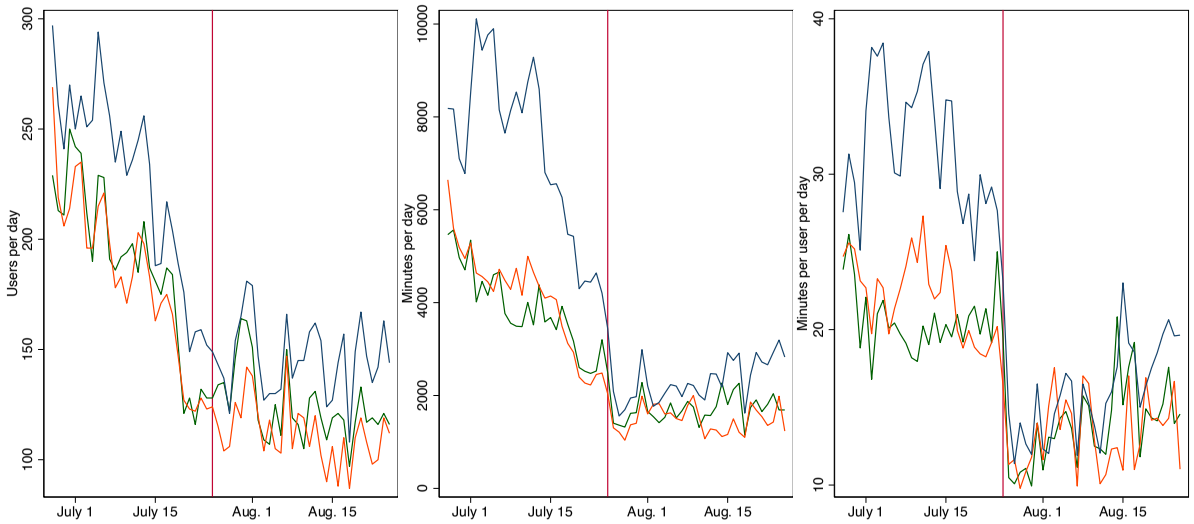
WHO. 2020. "Speeches of the Director General: Munich Security Conference." February, 15th.

WHO. 2021. "*Global Health Observatory*." Dataset accessed January 2021. https://www.who.int/data/gho/data/indicators/indicator-details/GHO/hospital-beds-(per-10-000-population).

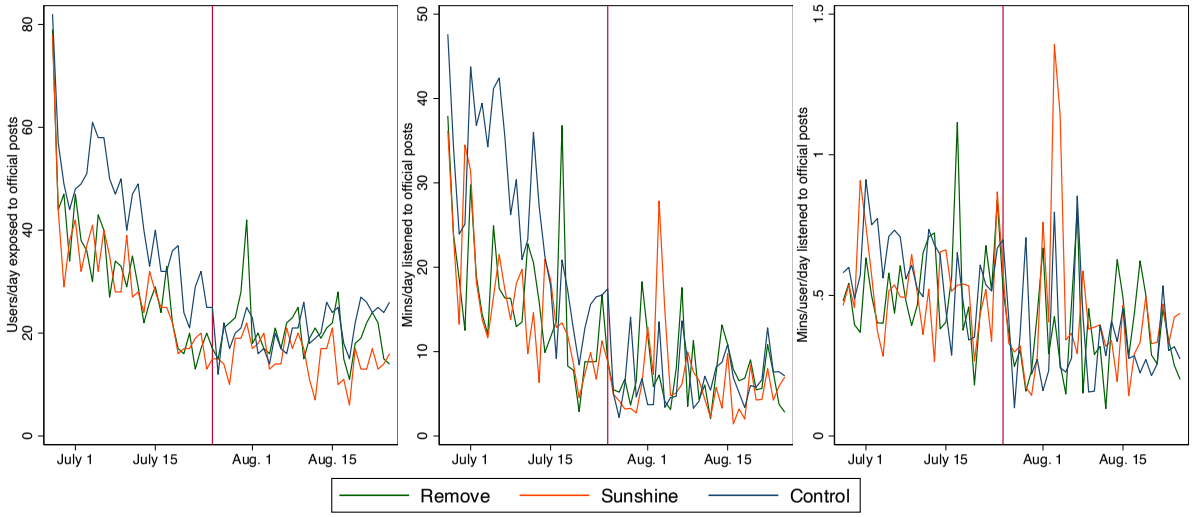
World Bank. 2006. *World Development Report 2007: Development and the Next Generation.* Washington, D.C.:World Bank.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov. 2020. "Political effects of the internet and social media." *Annual Review of Economics*, 12: 415–438.

Figure 1: User exposure to the platform during the experiment



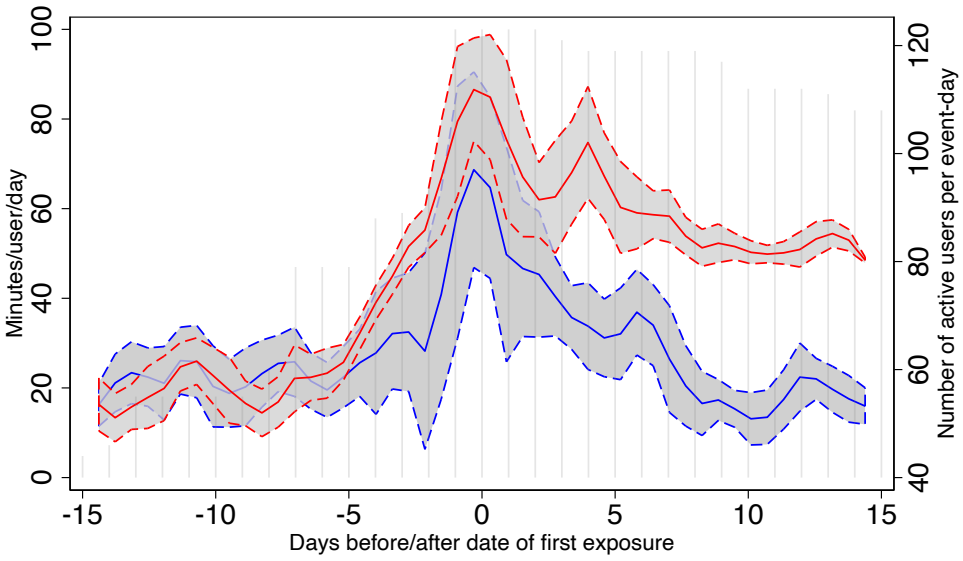
(a) Overall exposure to the platform



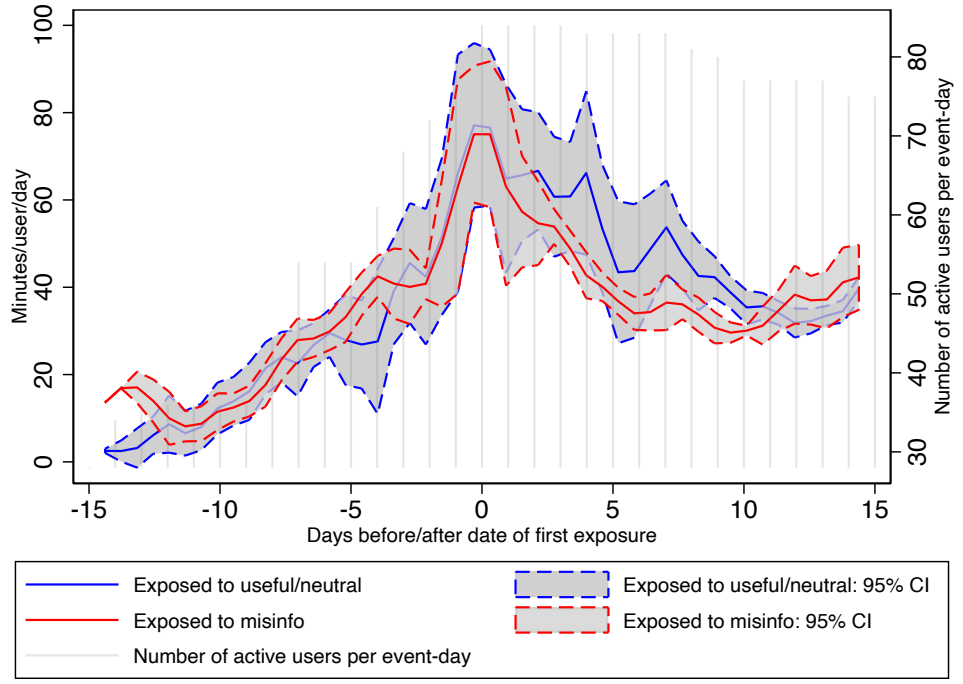
(b) Exposure to official information posts

Notes: The red vertical line represents the date that free access to the platform was limited for each user. In panel A: *Users per day* captures whether a user called in on that day and listened to at least one second of a post or comment. *Minutes per day* is the total minutes spent on the platform across all users per day. *Minutes per user per day* is the average minutes spent on the platform by users per day. In panel B: *Users/day exposed to official posts* captures whether a user listened at all to an official information post on that day. *Mins/day listened to official posts* is the total minutes spent listening to official posts across all users per day. *Mins/user/day listened to official posts* is the average minutes that a user spent listening to official posts on that day conditional on listening at all.

Figure 2: First misinformation exposure event study



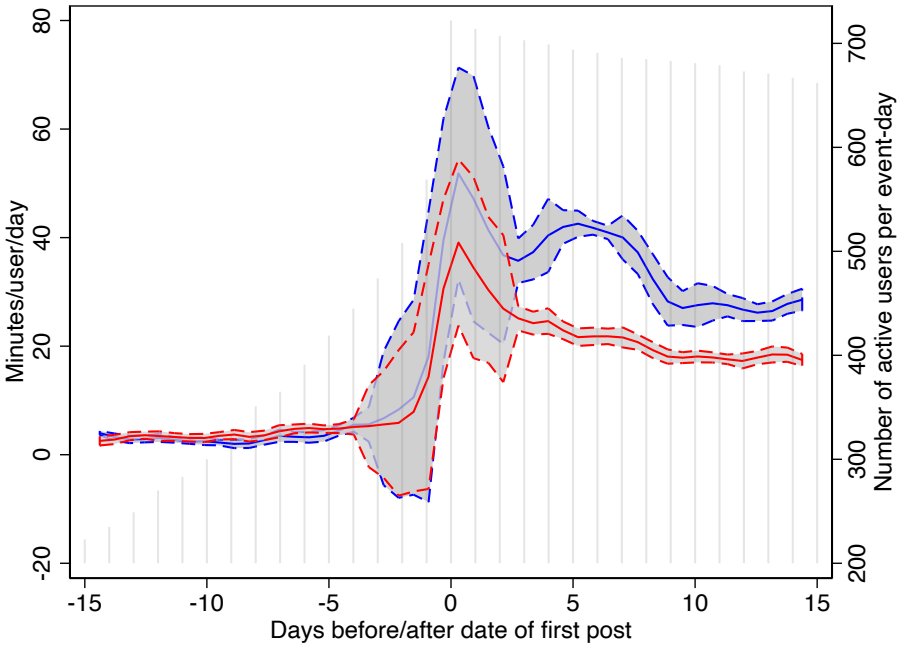
(a) Control users only



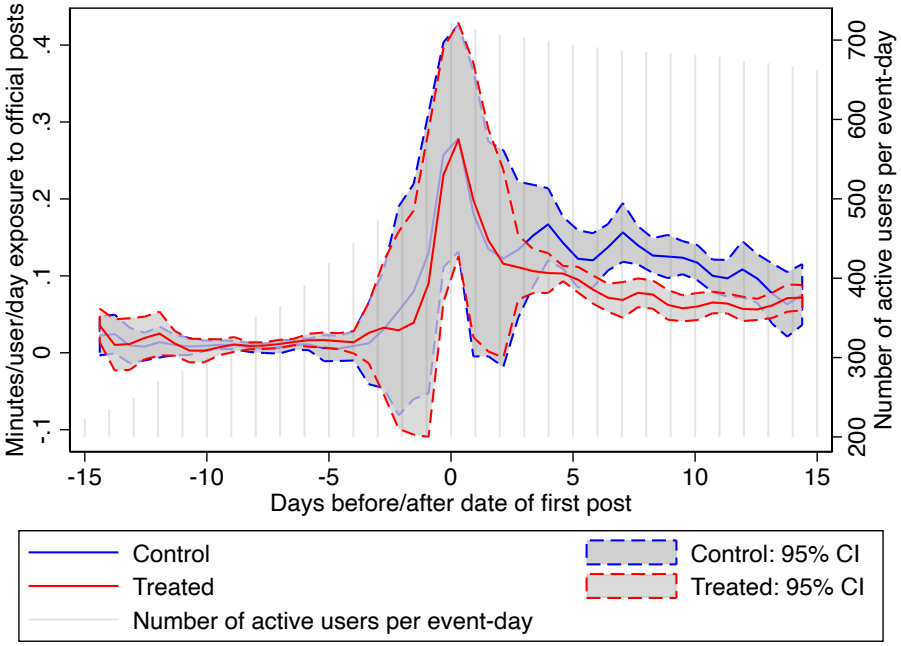
(b) Sunshine users only

Notes: This figure demonstrates an event study in which the treatment event at day zero is a user being exposed to a user-generated misinformation post for the first time. The matched counterfactual event is being exposed to a COVID-19-related but not misinformation post, selected to equal the number of misinformation posts via propensity score matching. The outcome measure in both panels is total minutes spent on the platform by user-event-day. Panel A considers only control users and Panel B considers only sunshine users. The sample is limited to users who posted at least once during the study period. Lines are local polynomial regressions with an epanechnikov kernel with bandwidth 1.

Figure 3: First post event study



(a) Overall exposure to the platform



(b) Exposure to official information posts

Notes: This figure demonstrates an event study in which the event at day zero is a user posting for the first time. The outcome measure in Panel A is total minutes spent on the platform by user-event-day. Panel B is minutes listened to an official information post on that user-event-day. The sample is limited to users who posted at least once during the study period. Lines are local polynomial regressions with an epanechnikov kernel with bandwidth 1.

Table 1: Main exposure and engagement outcomes for *official* information posts

	Minutes listened	Number of shares	Engagement index
Panel A: Original and pre-treatment referral users only			
Treatments combined			
Treated (=1)	-0.331**	-0.246	-0.043
	(0.168)	(0.156)	(0.028)
Treatments separated			
Remove (=1)	-0.305*	-0.292*	-0.025
	(0.181)	(0.167)	(0.032)
Sunshine (=1)	-0.357*	-0.199	-0.062**
	(0.188)	(0.177)	(0.031)
Remove = Sunshine?	0.731	0.518	0.177
Control mean	1.335	0.605	-0.000
# Clusters	1259	1259	1259
# Users	2077	2077	2077
Panel B: All post users			
Treatments combined			
Treated (=1)	-0.228**	-0.140	-0.053**
	(0.097)	(0.089)	(0.022)
Treatments separated			
Remove (=1)	-0.224**	-0.164*	-0.034
	(0.109)	(0.095)	(0.025)
Sunshine (=1)	-0.231**	-0.118	-0.071***
	(0.105)	(0.102)	(0.024)
Remove = Sunshine?	0.942	0.587	0.081
Control mean	0.883	0.356	-0.000
# Clusters	1408	1408	1408
# Users	3698	3698	3698

Notes: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. The unit of observation is the user. All outcome measures focus on official information posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or remove treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with clustered standard errors in parentheses. P-values are reported for the test that rejects Remove = Sunshine.

Table 2: Main exposure and engagement outcomes for *useful* information posts

	Minutes listened	Number of shares	Engagement index
Panel A: Original and pre-treatment referral users only			
Treatments combined			
Treated (=1)	-0.143**	-0.010	-0.022
	(0.064)	(0.007)	(0.031)
Treatments separated			
Remove (=1)	-0.139**	-0.006	-0.006
	(0.067)	(0.008)	(0.036)
Sunshine (=1)	-0.146**	-0.015**	-0.038
	(0.068)	(0.007)	(0.035)
Remove = Sunshine?	0.859	0.186	0.328
Control mean	0.350	0.022	-0.000
# Clusters	1259	1259	1259
# Users	2077	2077	2077
Panel B: All post users			
Treatments combined			
Treated (=1)	-0.084**	-0.009*	-0.023
	(0.037)	(0.005)	(0.023)
Treatments separated			
Remove (=1)	-0.073*	-0.004	-0.002
	(0.039)	(0.006)	(0.028)
Sunshine (=1)	-0.093**	-0.013***	-0.043*
	(0.040)	(0.005)	(0.025)
Remove = Sunshine?	0.446	0.045	0.115
Control mean	0.208	0.017	-0.000
# Clusters	1408	1408	1408
# Users	3698	3698	3698

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The unit of observation is the user. Useful information posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or remove treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with clustered standard errors in parentheses. P-values are reported for the test that rejects Remove = Sunshine.

Table 3: Main exposure and engagement outcomes for *misinformation* posts

	Minutes listened	Number of shares	Engagement index
Panel A: Original and pre-treatment referral users only			
Treatments combined			
Treated (=1)	-0.027	0.001	-0.086***
	(0.027)	(0.002)	(0.028)
Treatments separated			
Remove (=1)	-0.122***	-0.001	-0.115***
	(0.020)	(0.001)	(0.027)
Sunshine (=1)	0.068	0.003	-0.057*
	(0.043)	(0.003)	(0.034)
Remove = Sunshine?	0.000	0.082	0.007
Control mean	0.126	0.001	-0.000
# Clusters	1259	1259	1259
# Users	2077	2077	2077
Panel B: All post users			
Treatments combined			
Treated (=1)	-0.012	0.000	-0.068***
	(0.016)	(0.001)	(0.021)
Treatments separated			
Remove (=1)	-0.069***	-0.001	-0.089***
	(0.011)	(0.001)	(0.020)
Sunshine (=1)	0.040	0.002	-0.049*
	(0.026)	(0.002)	(0.025)
Remove = Sunshine?	0.000	0.088	0.013
Control mean	0.071	0.001	0.000
# Clusters	1408	1408	1408
# Users	3698	3698	3698

Notes: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. The unit of observation is the user. Misinformation posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or remove treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with clustered standard errors in parentheses. P-values are reported for the test that rejects Remove = Sunshine.

Table 4: User characteristics and attitudes

	Sample Means		
	Random	Most active	Exposed to misinformation
User characteristics			
Age	29.61	29.97	30.77
	(6.30)	(4.85)	(4.72)
Female (=1)	0.01	0.03	0.03
	(0.10)	(0.18)	(0.18)
Less than 8 years of education (=1)	0.19	0.10	0.17
	(0.40)	(0.31)	(0.38)
Less than 10 years of education (=1)	0.47	0.54	0.62
	(0.50)	(0.50)	(0.49)
Has a smartphone (=1)	0.91	0.83	0.80
	(0.28)	(0.38)	(0.40)
Uses WhatsApp at least once a day or more (=1)	0.91	0.78	0.76
	(0.28)	(0.42)	(0.43)
Perceptions of Baang content			
Trusts official more than users' COVID-19 posts (=1)	0.95	0.85	0.94
	(0.23)	(0.36)	(0.25)
Trust in official COVID-19 posts (1-5)	3.08	2.96	3.24
	(0.81)	(0.81)	(0.77)
Trust in users' COVID-19 posts (1-5)	2.23	1.92	2.01
	(0.91)	(0.80)	(0.78)
Prefers Baangs are moderated (=1)	0.99	1.00	1.00
	(0.10)	(0.00)	(0.00)
Prefers Baang team moderates (as opposed to users)	1.00	1.00	1.00
	(0.00)	(0.00)	(0.00)
Trust in sources of COVID-19 information (1-5)			
Government announcements	3.83	3.57	3.65
	(0.82)	(0.64)	(0.78)
Doctor	3.83	4.02	3.93
	(0.81)	(0.71)	(0.72)
Friends & family	3.64	3.75	3.87
	(0.65)	(0.55)	(0.66)
Local imam	3.09	3.16	3.23
	(0.80)	(0.73)	(0.81)
Social media	1.79	1.69	1.73
	(0.71)	(0.70)	(0.64)
Observations	94	87	86

Notes: *Random* is a random sample of users who ever called into the platform. The *Most active* sample are the users that comment the most times. The *Exposed to misinformation* sample are the users most exposed to misinformation. Standard deviations are reported in parentheses.

The spread of (mis)information: A social media experiment in Pakistan

Sarojini R. Hirshleifer, Mustafa Naseem, Agha Ali Raza, Arman Rezaee

Supplemental Appendix for Online Publication
September 19, 2023

SA1 Study Design

SA1.1 Categorizing and responding to COVID-19 content

SA1.1.1 Real-time moderation

Setting up a system to quickly identify and respond to misinformation on the platform required careful preparation. Before the experiment, we gathered and reviewed lists of COVID-19 myths and responses from a few reliable sources, particularly the WHO. We also catalogued types of misinformation that were circulating in Pakistan aside from the myths addressed by the WHO.

All user-generated content went through the same moderation process regardless of the condition assignment of the user who posted it. The moderation relied on a three-tiered system. First, experienced moderators who were specifically trained for the experiment reviewed all posts and comments and identified them as COVID-19-related or not. Next, senior research assistants reviewed the COVID-19 content and determined whether it was false.[61] To do so, they largely relied on the lists that were discussed and approved by the research team before the study. They also tagged misinformation according to broad categories such as denial of existence. If the research assistants could not make a determination, then a public health expert associated with the project made a final determination.

The categories of misinformation on the platform mostly were in line with myths that were widely identified and addressed by the WHO. Furthermore, most of them were unambiguously false. For a detailed categorization of misinformation on the platform, see Tables SA8 and SA9.

SA1.1.2 Additional content categorization

After the study was complete, research assistants conducted a full review of all user-generated COVID-19 posts and comments from the experiment.[62] In this review, the research team again lis-

[61] Tables SA8 and SA9 catalog all types of misinformation that appears on the platform during the study.
[62] Due to data limitations, we focus on posts in our analysis.

tened to and categorized all COVID-19 content. This second categorization of misinformation was then checked against the categorization that was used in the experiment. This process allowed us to verify that misinformation was accurately categorized, since initial categorization of the COVID-19 content was done under time pressure during the experiment. Less than 1% of content that had not been identified as misinformation during the experiment needed to be recategorized at this stage. We use the final categorization in conducting the analysis, which is why the amount of misinformation content in the remove treatment is close to but not exactly zero.

As part of this review, for the first time, we also further categorized user-generated COVID-19 content that was not misinformation as either neutral or *useful*. Specificly, we identified COVID-19 content as useful if it provided information about or recommended that people follow public health guidelines. Content that described specific instances of a user or someone they know getting or being treated for COVID-19 was also categorized as useful, since personal experiences can be helpful in providing evidence to users who might doubt the existence of COVID-19 or its seriousness. The remaining COVID-19 posts that were not useful or misinformation were categorized as neutral.

SA1.1.3 Crafting rebuttals

The rebuttals to address misinformation in the sunshine treatment were designed to provide relevant high-quality information with a quick turn around. Since the WHO website had, at the time, common myths about COVID-19 and responses to those myths, we relied on the WHO's myth responses to the extent possible. By drawing on information about the pandemic from credible health experts, we also drafted responses to additional, locally-relevant myths we had gathered by reviewing local social media (including Baang before the experiment). The tags that the research assistants assigned to misinformation facilitated in assigning rebuttals to posts. For example, once the research team assigned a tag to a post identifying it as misinformation about the origins of COVID-19, it was straightforward for them to assign a rebuttal that addressed the origins of the virus.

Since it was important to post user-generated content as soon as possible during the experiment, we prepared the rebuttals to misinformation posted in the sunshine treatment in advance to the extent possible. Thus, the majority of the rebuttals were recorded in advance of the experiment. This allowed moderators to post real-time responses to the misinformation that was posted on the platform. Still, there were some delays in posting misinformation in the sunshine treatment, since the average time to post misinformation was 554 minutes, while the average time to post COVID-19 content that was not misinformation was 270 minutes. COVID-19 misinformation, however, was only 1% of the content on the platform. Useful and neutral COVID-19 content was posted at

the same time in both the sunshine and remove treatments.

SA1.1.4 Defining study participants

A limitation of the randomization design is that we randomized people into treatments as soon as their first call was initiated. We did not realize that a large percentage of our sample would call the number for Baang, but never get past the main menu (41%). These may be accidental calls, or people who simply decide they are not interested in the platform after listening to the menu. The Baang main menu is identical regardless of condition assignment, however, thus individuals who never listen to any content beyond the main menu have not been exposed to any condition assignment. In fact, randomizing at the point of completing the menu would have been appropriate. Thus, we restrict the term *user* to refer to people who have been exposed to at least one second of content (i.e. posts or comments, whether from Baang or users) on the Baang platform during the experiment, and focus our analysis on these users. We find that our results are robust, however, to including individuals who never got past the menu. This is demonstrated in Section SA2.3 below.

Another consideration in understanding the randomization design is that we randomized based on phone number. We cannot rule out that a "user" in this case in fact represents multiple users who share a phone number. This would not be a direct threat to internal validity, however. Instead, if this is the case, we could think such a design as a randomization at the level of a group of users who share a phone number and who all are assigned to the same treatment. Less than 0.01% percent of calls come from a landline, but we cannot rule out sharing of cell phones. Still, we have no particular reason to believe that this situation is common in this setting.

Another possibility is that some users may have called in using multiple phone numbers, and thus were exposed to multiple treatments. Although this is possible, we believe that it would have been very difficult for users to become aware they were being assigned to a specific treatment, and systematically move from one treatment to another. We did not notify users that their posts were being moderated, so if they did experience a delay, it was likely to have been perceived as idiosyncratic. There were no announcements in the treatments, so for example, users in the remove treatment would not have known they were not hearing any misinformation. Thus, if they called in on a different number and were placed in a control and heard a piece of misinformation, it would not be obvious that they were being exposed to a different version of the platform. Finally, when some of the authors have experimented on this platform in the past, users have at times become aware of the experiment and posted about it on the platform. In this experiment, that did not happen. Thus, this type of imperfect compliance may exist, but it is likely to be largely random. If some people are partially randomly exposed to a different treatment than the one to which they are

initially assigned, that would tend to reduce the observed treatment effects.

SA2 Robustness analysis

This section examines the robustness of our main results to our study design. It examines pre-trends, the statistical significance of our main graphical treatment results, and robustness to potential spillovers, to the potential of non-random initial hang-ups, and to outliers. It also examines the effect of time to moderation on subsequent platform usage.

SA2.1 Significance of pre- and post-randomization trends

In Figure SA1, we present additional graphical analysis of our main hypothesis. These are local polynomial regressions with standard errors, which allows us to consider statistical significance. We combine treatments for clarity of exposition given the standard errors. This smoothed data is in contrast to Figure 1, which represents the raw data.

This analysis has a dual purpose. First, it confirms the statistical significance of the results in Figure 1. Second, it goes some way towards confirming the validity of the randomization. This second purpose has a key limitation. As discussed above, less than half of the users in the experiment called in before the experiment began. Thus, we are limited in our ability to test for the validity of the randomization, and cannot conduct a standard balance table on pre-treatment usage. Instead, in Figure SA1, we analysis similar to that in Figure 1, but focus on the time period from one month before the experiment began through the first month of the experiment. The blue line in each subpanel indicates the beginning of the experiment. There are no systematic differences across treatment and control in the pre-treatment period. Given that the available data during this period represents just a fraction of the full RCT sample, it is likely that these results would further converge if we had pre-treatment data for the full sample. Furthermore, there is a clearly exogenous break in the data as treatment begins, and as was evident in Figure 1, there are clear differences across treatment and control in five out the six subpanels.

SA2.2 Spillovers

Next, we consider the extent to which the design of the randomization accounted for referral networks. We do find that people are much more likely to share within their assigned condition than across conditions, suggesting that the original-referral user clusters do capture meaningful real-world networks (Figure SA2).

To confirm that cross-treatment sharing is not driving our results, we present our main results

while controlling for observed spillovers (Table SA2). These spillovers are potentially endogenous to treatment. This means that we cannot identify the effect of these spillovers on treatment. In this case, however, we are only interested in the controlled direct effect of the main treatment. That is, we only aim to confirm that our main treatment effects are not driven by the spillover channel. Thus, we implement the method proposed by Acharya, Blackwell and Sen (2016) to uncover the controlled direct effect. Specifically, we account for spillovers as measured through the number of shares a user received during the treatment period from users outside their treatment arm.[63]

The controlled direct effects accounting for spillovers are remarkably similar to the main results reported in Tables 1, 2, and 3, Panels A. For example, the coefficient on the combined treatments for the outcome of minutes listened is -0.331 in Table 1, and controlled direct effect is -0.344 for the same outcome when controlling for spillovers in Table SA2. We see a similar pattern when examining the results for the other two main outcomes (number of shares and the engagement index) and when examining the two treatments separately, as well as for both useful and misinformation content. We also see a similar pattern comparing Panels B of Tables 1, 2, and 3 with the results in Table SA5, which conducts this robustness check on the full experiment sample (-0.228 vs -0.231, etc.).

SA2.3 Hang-ups

As discussed in the randomization design Section SA1.1.4, some individuals were never exposed to any condition in the experiment because they hung up during the Baang main menu before reaching any Baang content. Since the main menu is the same for all conditions, these individuals have never been exposed to any condition. Thus, we do not include them in the main analysis. Note that this is not analogous to a typical partial compliance setting, because the individuals who hang-up during the main menu are evenly distributed across all three conditions. Thus, it would not be possible to use treatment assignment as an instrument, for example.

Instead, we examine the robustness of our main results to including these individuals (Table SA3). In this sample, 41% were never exposed to any condition. Thus, we would expect that the absolute magnitude of our coefficients would be reduced. If these hang-ups are the same across the three conditions, however, we would expect that the relative magnitude of our treatment effects and their significance to be robust. In practice, we do find that relative magnitude of treatment effects for this sample are effectively identical to those reported for the sample of users reported in Tables 1, 2, and 3, Panels A. The results also remain statistically significant. For example, the impact of being treated on minutes listened to official information posts is -0.246 (27%) compared to -0.331

[63] When examining the impact of the combined remove and sunshine treatments, a spillover identified as a share from someone in the control, while a spillover into the control is identified as share from either of the two treatments.

minutes (25%) in our main results. Similarly, for useful posts, the combined impact of treatment is -0.101 (42%) when including the hang-ups compared to -0.143 minutes (41%) for our main results. All four results are significant at the 5% level. We find the same pattern of robustness when considering our full experiment sample, comparing Tables 1, 2, and 3, Panels B to Table SA6.

SA2.4 Outliers

Finally, we consider the role of outliers in Table SA4. We find that our results are robust to accounting for them. Specifically, we winsorize our main outcomes, and replicate the main results (Tables 1, 2, and 3, Panels A). We winsorize at the 99th percentile, which is appropriate for our setting. As is the case with most social media platforms, a minority of people who ever use Baang during a two-month period (i.e. during the experiment) engage with the platform regularly. It is only within a limited subset of users, however, that treatment effects are potentially detectable. For example, only 46% of our sample listened to any official information posts during the experiment. Thus, in this case, winsorizing at the 99th percentile would actually remove over 2% rather than 1% of our potentially treated sample.

Our results are robust to removing these outliers. Although the absolute magnitude of the treatment effects are noticeably smaller, the relative magnitude of the treatment effects is close to our main results. The treatment effects are still meaningful in size, however, and significant at the 10% level. For example, the combined treatment effect on minutes listened to official information posts is -0.203 minutes (18%) compared to -0.331 minutes (25%) without winsorizing. Similarly, this pattern is reflected in the results for minutes listened to useful information posts, where the treatment effect is -0.081 minutes (30%) compared to -0.143 minutes (41%). This pattern generally holds across the other outcomes. It also holds in our full experimental sample, comparing Tables 1, 2, and 3, Panels B to Table SA7.

SA2.5 Time to moderation

In considering the delay mechanism, a question that arises is whether the length of time before a user's first post is moderated appears to be a significant factor in determining their subsequent usage of the platform. Thus, we consider that question more closely in this section. In Figure SA4, we examine the association between time to moderation and subsequent exposure to official posts using a non-parametric approach. Specifically, we plot the minutes to moderate a user's first post against that user's subsequent usage of the platform (Panel A) and exposure to official posts (Panel B).[64] In addition to plotting raw data points, we also plot local polynomial regression lines

[64] Both measures focus on total exposure after users post for the first time until the end of the experiment.

of best fit. We examine these results for treated and control users separately. Since all posts were moderated, time to moderation applies to both treatment and control users. Of course, control users did not experience their posts being moderated directly. As is the case for many commonly examined sources of treatment heterogeneity, we do not expect time to moderation to be randomly assigned. People who are moderated more quickly may be fundamentally different that those who are moderated more slowly, since how unusual a post is can be an important determinant of how quickly it is moderated. However, treatment assignment should not be correlated with time to moderation.[65]

We do not find that within-sample variation in time to moderation is a clear predictor of subsequent platform usage (Figure SA4). Specifically, it does not appear to be the case that delay mechanism is driven by the longest moderation times. Users exposed to either short or long delays seemed to have similar responses. Furthermore, the fact that the relationship between time to moderation and subsequent usage has the same trend across treatment and control reinforces that there are no differential treatment effects by time to moderation. This is not particularly surprising given that even short delays are likely to matter in the context of social media, where people may be looking for immediate gratification. A large percentage of the users who post and then check the main feed before their posts are moderated do so in first few minutes after posting.[66]

We note that the levels (as opposed to trends) across the treatment and control here are not statistically significantly different. The control is consistently higher than that of the treated, however, which is aligned with the event study delay mechanism. The standard error bars are too wide in this less-powered exercise to reject equality in exposure.

SA3 COVID-19 misinformation prevalence outside of Baang

We examine whether the misinformation statements posted on Baang during our study were prevalent in other traditional and social media sources at that same time. This is likely relevant to considering the potential impact of the sunshine treatment, which exposes users' to misinformation for the purpose of debunking it. There is some evidence that people do not fully update after being exposed to debunking. If people have already been exposed to some piece of misinformation outside the platform, however, then the benefit from debunking this misinformation is likely to outweigh the presumably minimal marginal cost of exposing people again to misinformation they have already encountered. This is likely the theory behind the many debunking websites that arose

[65] Note that control arm posts went through an identical moderation process as treatment arm posts before going live in those treatment arms, so time to moderation is measured equally across arms.

[66] In addition, we do not observe delays of zero minutes; the 5th percentile is a roughly 3 minute delay.

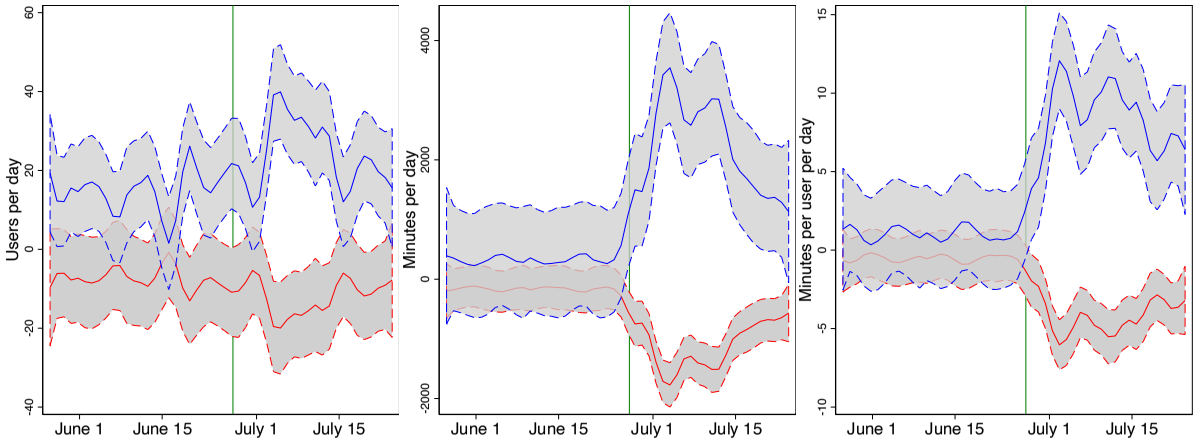
in the early days of the pandemic, including the widely disseminated debunking website produced by the WHO.

To gauge misinformation prevalence outside of Baang, we conducted a systematic review of a number of information sources in Pakistan. First, we reviewed social media, namely Twitter, which we consider a direct source of misinformation. Second, we reviewed sources that were attempting to debunk myths, namely government and policy organizations, on the theory that these sources had likely directly encountered these myths, which is why they had been chosen for debunking. Third, we reviewed traditional media, which was sometimes a direct source of misinformation and sometimes and indirect source seeking to debunk it. All reviews were conducted manually by trained research assistants for the entire study period. For Twitter, this entailed manually searching all Twitter trends from the period.[67] For government and policy organizations, research assistants reviewed websites and Twitter feeds for the Department of Health (including the Department's official COVID-19 site), NIH, Pakistan, Office of the Prime Minister, and several policy organizations with pages on myths in Pakistan such as the Friedrich Naumann Foundation. Traditional media included Dawn, Samaa News, and ARY News. Research assistants identified posts on all of these sources related to COVID-19 and manually categorized any posts into those same categories our moderators used during the study (in fact the research assistants were our moderators). Any posts identified as misinformation were then associated with misinformation posts from our platform as relevant. We then summed the number of times each piece of misinformation was found through this search in Tables SA8 and SA9.

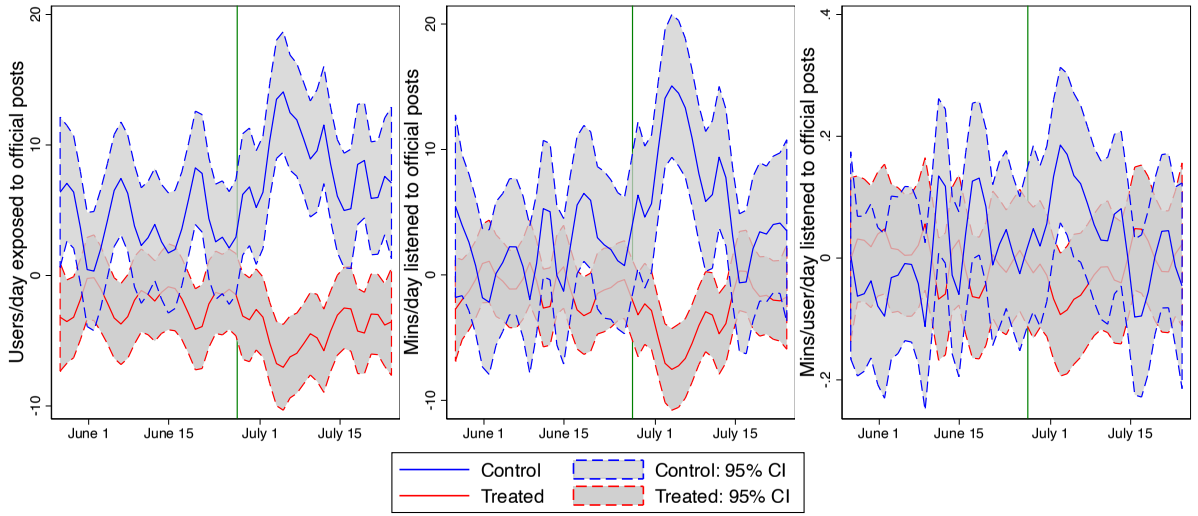
Of 42 unique myths, 26% were found on Twitter, 79% were found on government and policy pages/posts, and 62% were found on traditional media pages/posts. As some of mentions of these types of misinformation in traditional media were direct sources and some indirect, we were able to identify between 26 to 69% through direct sources. Still, it is likely that misinformation was chosen for debunking in indirect sources because those organizations had in fact encountered the misinformation directly. Thus, many if not most of the pieces of misinformation on Baang were being repeated in Pakistan outside of Baang. This suggests that users may have already encountered much of misinformation on the platform from other sources.

[67] Thus, we are only counting misinformation that became major topics on Twitter, which means that this does not fully capture all misinformation on the platform. It is likely then that at least some of the misinformation found on our platform that is not identified in this measure was also on Twitter.

Figure SA1: User exposure to the platform pre-trends



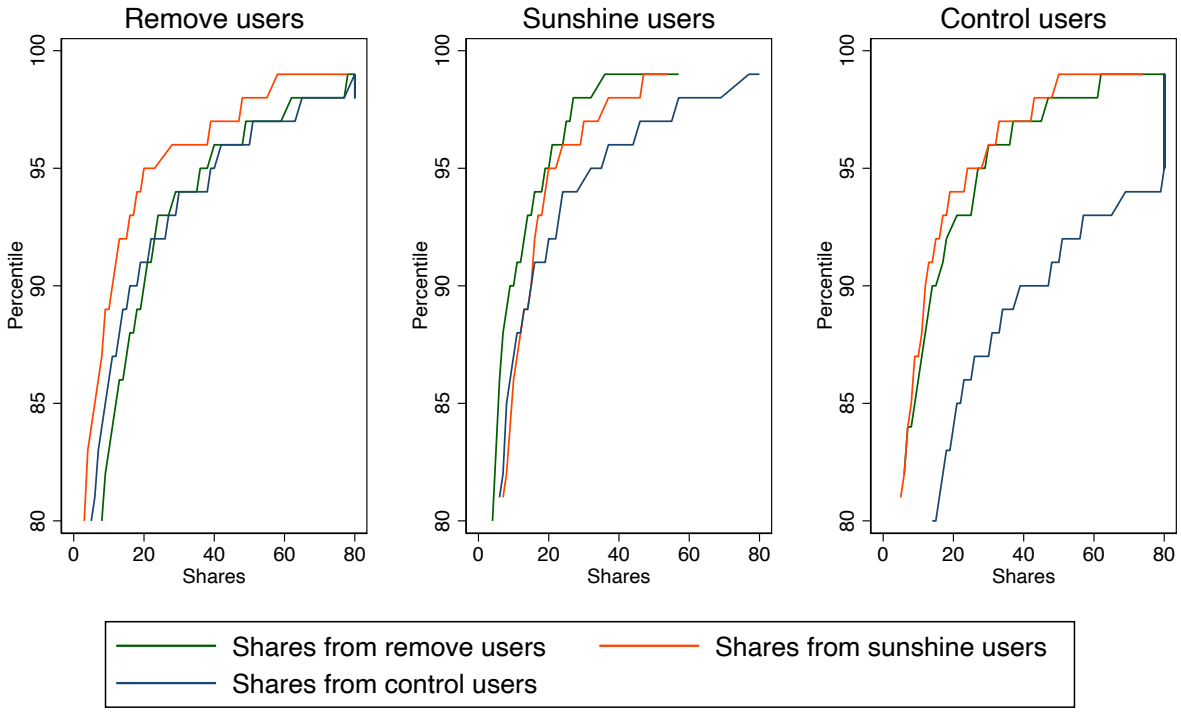
(a) Overall exposure to the platform



(b) Exposure to official information posts

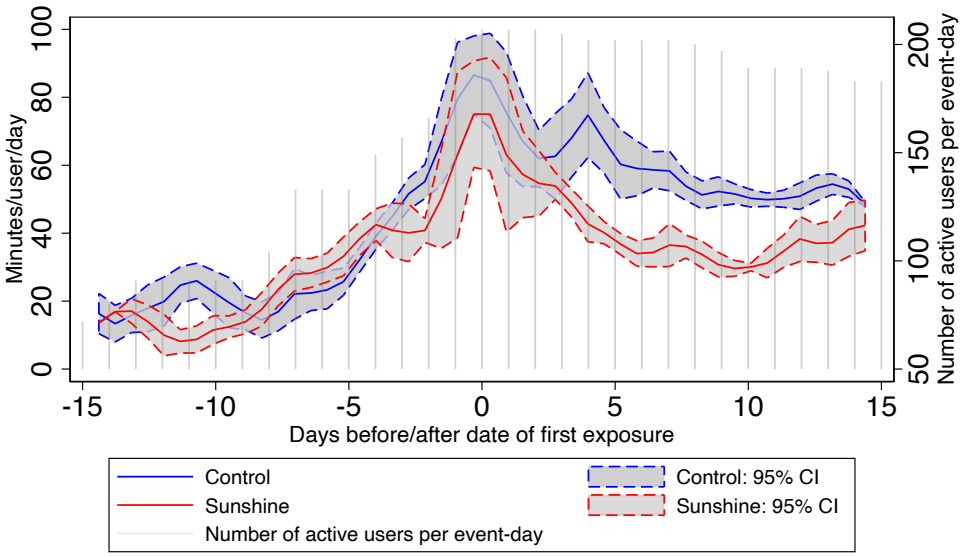
Notes: The green vertical line represents the date that treatment began. In panel A: *Users per day* captures whether a user called in on that day and listened to at least one second of a post or comment. *Minutes per day* is the total minutes spent on the platform across all users per day. *Minutes per user per day* is the average minutes spent on the platform by users per day. In panel B: *Users/day exposed to official posts* captures whether a user listened at all to an official information post on that day. *Mins/day listened to official posts* is the total minutes spent listening to official posts across all users per day. *Mins/user/day listened to official posts* is the average minutes that a user spent listening to official posts on that day conditional on listening at all. All data is residualized to remove date fixed effects. Lines are local polynomial regressions with an epanechnikov kernel with bandwidth 1.

Figure SA2: Shares within and across treatment arms



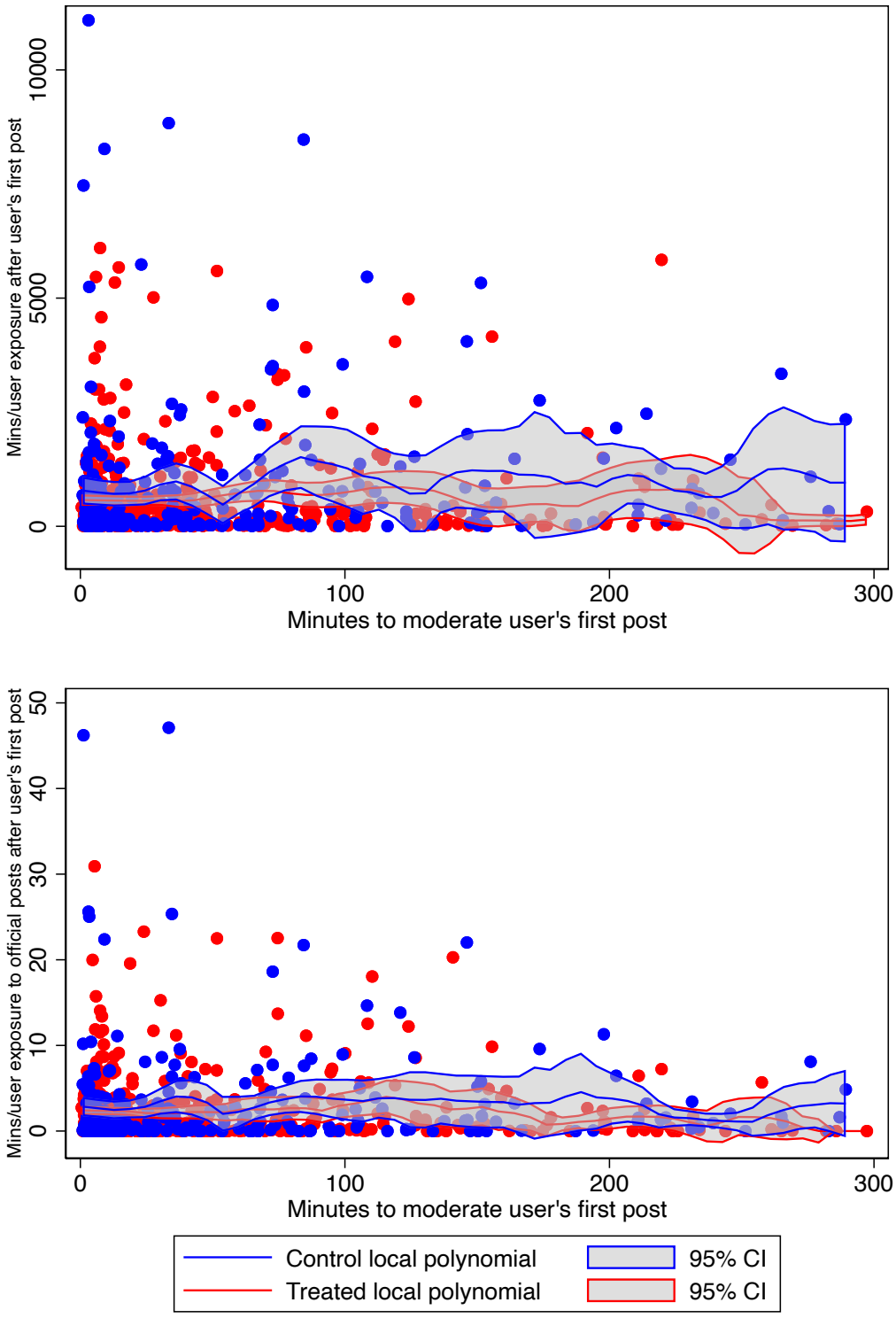
Notes: This figure displays the number of shares each user received during the experiment. The panels are separated by the user's treatment assignment, and the three lines within each panel represents the treatment assignment from which a given share came from. Number of shares winsorized at the 99th percentile.

Figure SA3: First misinformation exposure event study, control vs sunshine



Notes: This figure demonstrates an event study in which the event at day zero is a user being exposed to a user-generated misinformation post for the first time. The outcome measure is total minutes spent on the platform by user-event-day. The sample is limited to users who posted at least once during the study period. Lines are local polynomial regressions with an epanechnikov kernel with bandwidth 1.

Figure SA4: Time to moderation and exposure



Notes: This figure displays the total (panel A) and official (panel B) minutes users listened to the platform after each users first post, by the amount of time it took to moderate that first post. Sample is limited to users who ever post during the study period. Lines are local polynomial regressions with an epanechnikov kernel with bandwidth 1.

Table SA1: Measures of engagement for COVID-19 posts

	Listens	Number of shares	Engagement index
Official	653.2***	62.5***	13.0***
	(15.073)	(1.817)	(2.043)
Useful	3.2**	0.2**	-0.0
	(1.366)	(0.071)	(0.186)
Neutral	7.2***	0.4***	0.5**
	(1.640)	(0.078)	(0.214)
Constant	11.0***	0.0	0.0
	(0.913)	(0.038)	(0.167)
Official = Useful?	0.000	0.000	0.000
Official = Neutral?	0.000	0.000	0.000
Useful = Neutral?	0.019	0.007	0.001
# Posts	340	340	340

Notes: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. Limits to engagement by original and pre-treatment referral users in the control arm during the post period. *Official* is an indicator for an official information post. *Useful* is an indicator for a user-generated post that contains useful information about COVID-19. *Neutral* is an indicator for a user-generated COVID-19 post that does not contain useful or false information. The excluded category is user-generated misinformation posts, and the coefficients on the constant can be interpreted as the means of the excluded category. *Listens* is ever begins listening to a post. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of misinformation posts. Reports OLS regressions with standard errors in parentheses.

Table SA2: Main exposure and engagement outcomes controlling for spillovers

	Minutes listened	Number of shares	Engagement index
Panel A Outcome: Official *information posts*			
Treatments combined			
Treated (=1)	-0.344**	-0.252	-0.045
	(0.156)	(0.156)	(0.028)
Treatments separated			
Remove (=1)	-0.361**	-0.319*	-0.031
	(0.173)	(0.166)	(0.031)
Sunshine (=1)	-0.328*	-0.184	-0.059*
	(0.170)	(0.177)	(0.030)
Remove = Sunshine	0.817	0.336	0.288
Control mean	1.335	0.605	-0.000
Panel B Outcome: Useful *information posts*			
Treatments combined			
Treated (=1)	-0.146**	-0.010	-0.024
	(0.061)	(0.007)	(0.029)
Treatments separated			
Remove (=1)	-0.154**	-0.007	-0.016
	(0.065)	(0.008)	(0.035)
Sunshine (=1)	-0.138**	-0.014**	-0.032
	(0.064)	(0.007)	(0.032)
Remove = Sunshine	0.690	0.270	0.627
Control mean	0.350	0.022	-0.000
Panel C Outcome: Misinformation *posts*			
Treatments combined			
Treated (=1)	-0.029	0.001	-0.088***
	(0.027)	(0.002)	(0.027)
Treatments separated			
Remove (=1)	-0.127***	-0.002	-0.122***
	(0.020)	(0.002)	(0.026)
Sunshine (=1)	0.071*	0.003	-0.054*
	(0.042)	(0.003)	(0.031)
Remove = Sunshine	0.000	0.065	0.002
Control mean 0.126	0.001	-0.000	
# Clusters	1259	1259	1259
# Users	2077	2077	2077

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. Regressions control for spillovers from other treatment arms during the study period. The unit of observation is the user. Limited to original and pre-treatment referral users. Useful information and misinformation posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or remove treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with clustered standard errors in parentheses. P-values are reported for the test that rejects Remove = Sunshine.

Table SA3: Main exposure and engagement outcomes including hang-ups

	Minutes listened	Number of shares	Engagement index
Panel A outcome: Official *information posts*			
Treatments combined			
Treated (=1)	-0.246**	-0.175*	-0.040*
	(0.105)	(0.104)	(0.023)
Treatments separated			
Remove (=1)	-0.261**	-0.219**	-0.031
	(0.116)	(0.110)	(0.025)
Sunshine (=1)	-0.231**	-0.129	-0.050**
	(0.113)	(0.117)	(0.025)
Remove = Sunshine?	0.750	0.320	0.348
Control mean	0.897	0.408	-0.000
Panel B outcome: Useful *information posts*			
Treatments combined			
Treated (=1)	-0.101**	-0.007	-0.021
	(0.041)	(0.004)	(0.024)
Treatments separated			
Remove (=1)	-0.107**	-0.005	-0.015
	(0.043)	(0.005)	(0.028)
Sunshine (=1)	-0.095**	-0.010**	-0.027
	(0.043)	(0.005)	(0.026)
Remove = Sunshine?	0.655	0.277	0.638
Control mean	0.235	0.015	0.000
Panel C outcome: Misinformation *posts*			
Treatments combined			
Treated (=1)	-0.021	0.001	-0.072***
	(0.018)	(0.001)	(0.021)
Treatments separated			
Remove (=1)	-0.085***	-0.001	-0.098***
	(0.013)	(0.001)	(0.021)
Sunshine (=1)	0.045	0.002	-0.045*
	(0.028)	(0.002)	(0.025)
Remove = Sunshine?	0.000	0.066	0.003
Control mean	0.084	0.001	-0.000
# Clusters	1825	1825	1825
# Users	3159	3159	3159

Notes: $*p < 0.1$, $**p < 0.05$, $***p < 0.01$. Unit of observation is the user. Limited to original and pre-treatment referral users. Useful information and misinformation posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or remove treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with clustered standard errors in parentheses. P-values are reported for the test that rejects Remove = Sunshine.

Table SA4: Main exposure and engagement outcomes winsorized at 99th percentile

	Minutes listened	Number of shares	Engagement index
Panel A Outcome: Official *information posts*			
Treatments combined			
Treated (=1)	-0.203*	-0.115*	-0.024
	(0.123)	(0.062)	(0.032)
Treatments separated			
Remove (=1)	-0.174	-0.151**	0.004
	(0.136)	(0.070)	(0.037)
Sunshine (=1)	-0.231	-0.079	-0.052
	(0.144)	(0.072)	(0.036)
Remove = Sunshine	0.667	0.286	0.143
Control mean	1.160	0.360	-0.000
Panel B Outcome: Useful *information posts*			
Treatments combined			
Treated (=1)	-0.081**	-0.009	-0.025
	(0.039)	(0.006)	(0.034)
Treatments separated			
Remove (=1)	-0.076*	-0.005	-0.014
	(0.042)	(0.007)	(0.038)
Sunshine (=1)	-0.086**	-0.013**	-0.035
	(0.044)	(0.006)	(0.040)
Remove = Sunshine	0.784	0.113	0.588
Control mean	0.274	0.019	-0.000
Panel C Outcome: Misinformation *posts*			
Treatments combined			
Treated (=1)	-0.046**	0.001	-0.092***
	(0.020)	(0.002)	(0.030)
Treatments separated			
Remove (=1)	-0.110***	-0.001	-0.130***
	(0.017)	(0.001)	(0.029)
Sunshine (=1)	0.019	0.003	-0.053
	(0.027)	(0.003)	(0.036)
Remove = Sunshine	0.000	0.082	0.003
Control mean	0.114	0.001	0.000
# Clusters	1259	1259	1259
# Users	2077	2077	2077

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The unit of observation is the user. Limited to original and pre-treatment referral users. Useful information and misinformation posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or remove treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with clustered standard errors in parentheses. P-values are reported for the test that rejects Remove = Sunshine. Outcomes are winsorized at the 99th percentile if that value is not 0.

Table SA5: Main exposure and engagement outcomes controlling for spillovers, full experiment sample

	Minutes listened	Number of shares	Engagement index
Panel A Outcome: Official *information posts*			
Treatments combined			
Treated (=1)	-0.231**	-0.142	-0.054**
	(0.091)	(0.089)	(0.021)
Treatments separated			
Remove (=1)	-0.268***	-0.185*	-0.040*
	(0.102)	(0.095)	(0.024)
Sunshine (=1)	-0.197**	-0.102	-0.067***
	(0.097)	(0.100)	(0.022)
Remove = Sunshine	0.387	0.304	0.150
Control mean	0.883	0.356	-0.000
Panel B Outcome: Useful *information posts*			
Treatments combined			
Treated (=1)	-0.085**	-0.009*	-0.024
	(0.035)	(0.005)	(0.022)
Treatments separated			
Remove (=1)	-0.086**	-0.005	-0.012
	(0.037)	(0.006)	(0.027)
Sunshine (=1)	-0.084**	-0.013***	-0.035
	(0.036)	(0.005)	(0.023)
Remove = Sunshine	0.947	0.086	0.335
Control mean	0.208	0.017	-0.000
Panel C Outcome: Misinformation *posts*			
Treatments combined			
Treated (=1)	-0.013	0.000	-0.068***
	(0.015)	(0.001)	(0.019)
Treatments separated			
Remove (=1)	-0.073***	-0.001	-0.095***
	(0.011)	(0.001)	(0.019)
Sunshine (=1)	0.043*	0.002	-0.044*
	(0.025)	(0.002)	(0.022)
Remove = Sunshine	0.000	0.061	0.002
Control mean	0.071	0.001	0.000
# Clusters	1408	1408	1408
# Users	3698	3698	3698

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. Regressions control for spillovers from other treatment arms during the study period. The unit of observation is the user. Includes all users that called in and made it past the menu during the experiment. Useful information and misinformation posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the su. nshine or remove treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with clustered standard errors in parentheses. P-values are reported for the test that rejects Remove = Sunshine.

Table SA6: Main exposure and engagement outcomes including hang-ups, full experiment sample

	Minutes listened	Number of shares	Engagement index
Panel A outcome: Official *information posts*			
Treatments combined			
Treated (=1)	-0.142**	-0.086*	-0.042**
	(0.055)	(0.051)	(0.017)
Treatments separated			
Remove (=1)	-0.155**	-0.111**	-0.028
	(0.064)	(0.056)	(0.019)
Sunshine (=1)	-0.131**	-0.066	-0.054***
	(0.057)	(0.058)	(0.018)
Remove = Sunshine?	0.624	0.347	0.094
Control mean	0.527	0.213	-0.000
Panel B outcome: Useful *information posts*			
Treatments combined			
Treated (=1)	-0.051**	-0.005*	-0.019
	(0.021)	(0.003)	(0.017)
Treatments separated			
Remove (=1)	-0.050**	-0.003	-0.008
	(0.023)	(0.004)	(0.021)
Sunshine (=1)	-0.052**	-0.007***	-0.028
	(0.022)	(0.003)	(0.017)
Remove = Sunshine?	0.915	0.081	0.304
Control mean	0.124	0.010	-0.000
Panel C outcome: Misinformation *posts*			
Treatments combined			
Treated (=1)	-0.008	0.000	-0.053***
	(0.009)	(0.001)	(0.015)
Treatments separated			
Remove (=1)	-0.045***	-0.001	-0.074***
	(0.007)	(0.001)	(0.015)
Sunshine (=1)	0.023	0.001	-0.035**
	(0.016)	(0.001)	(0.017)
Remove = Sunshine?	0.000	0.058	0.001
Control mean	0.043	0.000	-0.000
# Clusters	2002	2002	2002
# Users	6239	6239	6239

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. Unit of observation is the user. Includes all users that called in during the experiment regardless of whether they ever made it past the menu. Useful information and misinformation posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or remove treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with clustered standard errors in parentheses. P-values are reported for the test that rejects Remove = Sunshine.

Table SA7: Main exposure and engagement outcomes winsorized at 99th percentile, full experiment sample

	Minutes listened	Number of shares	Engagement index
Panel A Outcome: Official *information posts*			
Treatments combined			
Treated (=1)	-0.141**	-0.054**	-0.041*
	(0.068)	(0.025)	(0.024)
Treatments separated			
Remove (=1)	-0.138*	-0.073***	-0.013
	(0.080)	(0.027)	(0.029)
Sunshine (=1)	-0.144*	-0.037	-0.067**
	(0.076)	(0.031)	(0.028)
Remove = Sunshine	0.938	0.216	0.075
Control mean	0.751	0.175	0.000
Panel B Outcome: Useful *information posts*			
Treatments combined			
Treated (=1)	-0.039**	-0.009*	-0.023
	(0.020)	(0.005)	(0.027)
Treatments separated			
Remove (=1)	-0.029	-0.004	0.009
	(0.021)	(0.006)	(0.032)
Sunshine (=1)	-0.049**	-0.013***	-0.053*
	(0.023)	(0.005)	(0.029)
Remove = Sunshine	0.326	0.045	0.056
Control mean	0.148	0.017	-0.000
Panel C Outcome: Misinformation *posts*			
Treatments combined			
Treated (=1)	-0.025***	0.000	-0.070***
	(0.009)	(0.001)	(0.022)
Treatments separated			
Remove (=1)	-0.056***	-0.001	-0.097***
	(0.008)	(0.001)	(0.020)
Sunshine (=1)	0.005	0.002	-0.046*
	(0.014)	(0.002)	(0.026)
Remove = Sunshine	0.000	0.088	0.006
Control mean	0.058	0.001	0.000
# Clusters	1408	1408	1408
# Users	3698	3698	3698

Notes: *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The unit of observation is the user. Includes all users that called in and made it past the menu during the experiment. Useful information and misinformation posts are user-generated. All outcome measures focus on posts about COVID-19. *Treated* is an indicator for being assigned to either the sunshine or remove treatments. *Engagement index* is a z-score average of comments, likes, and dislikes, with each engagement normalized relative to the mean and standard deviation of control users' posts. Reports OLS regressions with clustered standard errors in parentheses. P-values are reported for the test that rejects Remove = Sunshine. Outcomes are winsorized at the 99th percentile if that value is not 0.

Table SA8: COVID-19 Misinformation Prevalence by Source I

Misinformation Statement	*Source*			
	Soc. Media	Gov't & Policy	Trad. Media	Total
COVID-19 has a cure	1	5	2	8
Herbs are a cure	1	4	2	7
No need to maintain precautions	0	4	2	6
COVID-19 does not effect people in a certain geographic areas	1	2	2	5
COVID-19 does not exist	0	5	0	5
COVID-19 is a global conspiracy	0	3	2	5
COVID-19 is a punishment by God	0	2	3	5
COVID-19 is a scheme by Bill Gates to insert microchips in people	0	4	1	5
Religious practices help protect against COVID-19	0	2	3	5
Suggestion of false cures such as antibiotics	1	2	2	5
Virus only spreads in certain weather	1	3	1	5
COVID-19 is caused by a disregard of religious values	0	2	2	4
Official sources are misreporting	1	2	1	4
Spread of COVID-19 misreported/exaggerated	0	3	1	4
Warm water is a cure	1	2	1	4
COVID-19 does not affect Muslims	0	1	2	3
COVID-19 is a conspiracy against Muslims	0	1	2	3
COVID-19 is a misunderstanding on the Govt. of Pakistan's part	0	3	0	3
COVID-19 is not to be taken seriously	0	2	1	3
COVID-19-related death toll is exaggerated/made-up	0	2	1	3
Onions are a cure	1	2	0	3
People are falsely being labelled COVID-19 positive (under a conspiracy)	0	1	2	3
Poisonous injections in the name of COVID-19	0	1	2	3

Notes: See Section SA3 for a detailed explanation of this table.

Table SA9: COVID-19 Misinformation Prevalence by Source II

Misinformation Statement	*Source* Soc. Media	Gov't & Policy	Trad. Media	Total
COVID-19 is a conspiracy of doctors	0	0	2	2
COVID-19 is a conspiracy of the Govt. of Pakistan	0	1	1	2
COVID-19 is an old disease	0	2	0	2
Dehydration causes COVID-19	1	0	1	2
Okay to hug	0	2	0	2
Okay to shake hands	0	2	0	2
Pandemic is over	1	1	0	2
Saline water is a cure	1	0	1	2
Washing masks with dettol soap makes them re-usable	0	2	0	2
COVID-19 does not spread during Summers/in heat	0	1	0	1
COVID-19 is a conspiracy to get aid/get foreign debt forgiven	0	0	1	1
COVID-19 is a scheme to harvest organs	0	1	0	1
Steam as a cure	0	1	0	1
There is no COVID-19 in Pakistan/an area of Pakistan	0	1	0	1
We should stop talking about COVID-19	0	0	1	1
COVID-19 is a conspiracy against the poor	0	0	0	0
COVID-19 is a money making scheme	0	0	0	0
Tea/similar as a cure	0	0	0	0
Using Baang in a particular way causes/prevents COVID-19	0	0	0	0

Notes: See Section SA3 for a detailed explanation of this table.