DESIGNING EFFECTIVE CELEBRITY PUBLIC HEALTH MESSAGING: RESULTS FROM A NATIONWIDE TWITTER EXPERIMENT IN INDONESIA

VIVI ALATAS[⋆], ARUN G. CHANDRASEKHAR[†], MARKUS MOBIUS[§], BENJAMIN A. OLKEN[‡], AND CINDY PALADINES[⋆⋆]

Abstract. While celebrity messaging is thought to be important in the design of public health campaigns, there is little work (a) decomposing the extent to which celebrity endorsement relies on directly authoring the content as opposed to passing on messages of others, (b) measuring how much explicitly citing public health authorities matters, and (c) examining whether these online messages are meaningfully received and digested offline. We address these three questions by conducting a large-scale randomized trial in Indonesia in 2015-2016, in which 46 high profile celebrities agreed to tweet or retweet messages promoting immunization. The decomposition shows that public health messages are 72% more likely to be passed on when a celebrity is involved, but importantly, almost all of the effect (79%) comes only when the celebrity authors the message themselves. Explicitly citing an external medical authority reduces the passing-on rate by 27%. The results suggest that celebrities have an outsize influence in shaping public opinion on public health, particularly when they speak in their own voice.

Date: June 26, 2020.

All authors contributed equally. We thank Ran Abramitsky, Marcella Alsan, Nancy Baym, Emily Breza, Leo Bursztyn, Rebecca Diamond, Dean Eckles, Paul Goldsmith-Pinkham, Ben Golub, Mary Gray, Rema Hanna, Johannes Haushofer, Matt Jackson, Tyler McCormick, Matthew Wai-Poi, Alex Wolitsky, and participants at various seminars for helpful discussions. Aaron Kaye, Nurzanty Khadijah, Devika Lakhote, Eva Lyubich, Sinead Maguire, Lina Marliani, Sebastian Steffen, Vincent Tanutama provided excellent research assistance. We thank Nila Moeloek, then Indonesian Special Envoy for Sustainable Development Goals, Diah Saminarsih, and their team for providing support for this project. This study was approved by IRBs at MIT (Protocol #1406006433) and Stanford (Protocol #31451), and registered in the AEA Social Science Registry (AEARCTR-0000757). Funding for this project came from the Australian Government Department of Foreign Affairs and Trade. The views expressed here are those of the authors only and do not represent those of any of the institutions or individuals acknowledged here.

[⋆] Asa Kreativita.
[†] Department of Economics, Stanford University; NBER; J-PAL.
[§] Microsoft Research, New England.
[‡] Department of Economics, MIT; NBER; J-PAL.
[⋆⋆] World Bank.

1. Introduction

Social media has allowed celebrities to take an increasing role in social discourse. With millions of followers, celebrities have a channel to spread messages on many issues, including some far removed from their original reason for fame. Their participation in ongoing discussions can make issues prominent and shape the zeitgeist.

Examples abound. #BlackLivesMatter, a campaign against racial injustice, is the most-used social issue Twitter hashtag ever, with 41 million uses as of September 2018, with hundreds of thousands more in June surrounding protests sparked by the death of George Floyd. In public health, the #IceBucketChallenge, promoting awareness of Lou Gehrig's disease, became the sixth most-used social issue hashtag ever following participation by celebrities, from Oprah to Bill Gates. Each of these campaigns was initiated by a less-well-known activist, but made prominent in part through celebrity participation.

Using celebrities effectively for a public health campaign, however, depends on what features of the celebrity messaging spur diffusion. Beyond the direct *reach* that celebrities have – they have numerous followers – there are several dimensions that one must consider for policy design, and measuring the effect of each is crucial. First, how much does celebrity *endorsement*, meaning involvement with the messaging beyond the simple fact that they have numerous followers, matter? Second, how much of an endorsement premium comes from direct *authorship effect* rather than relaying the message of others including public health authorities. Third, how much does inclusion of *credible sourcing* matter, particularly when celebrities are speaking on a topic removed from their core area of expertise such as public health?

While it is clear, ex-ante, that celebrity endorsement is likely to matter, an optimal public health design crucially depends on the answers to (1)-(3) above. For instance, if endorsement matters, but authorship has limited value, and followers find that passing information from credible medical authorities is convincing, then a public health campaign should simply have celebrities pass on messages from public health authorities to their followers. In contrast, if messages only have value if celebrities write in their own voice, perhaps to the extent that citing public health authorities discourages information diffusion, then optimal design looks very different.

Measuring and decomposing the endorsement, authorship, and credible sourcing effects, as well as measuring whether any of these campaigns have offline effects, is challenging. First, celebrities' decisions about whether, and what, to say publicly are choice variables and influenced by the general information environment into which they are speaking. People also consume information from so many sources that in most contexts it is also nearly impossible to isolate the response to a particular piece of information or exposure to a particular celebrity, either online or offline. And, in general, a given action by a celebrity bundles reach and endorsement, making it hard to disentangle precisely which messages have an impact.

To study these issues, we conducted an experiment through a nationwide immunization campaign on Twitter from 2015-2016 in Indonesia, in collaboration with the Indonesian Government's Special Ambassador to the United Nations for Millennium Development Goals. Working with the Special Ambassador, we recruited 46 high-profile celebrities and organizations, with a total of over 11 million followers, each of whom gave us access to send up to 33 tweets or retweets promoting immunization from their accounts. The content and timing of these tweets was randomly chosen from a bank approved by the Indonesian Ministry of Health, all of which featured a campaign hashtag #AyoImunisasi ("Let's Immunize").

The experiment randomly varied the tweets along three key dimensions: (1) Did the celebrity / organization send the tweet directly, or did they retweet a message (drawn randomly from the same tweet library) sent by us from an ordinary (non-celebrity) user's account?; (2) Did the tweet explicitly cite a public health source?; and (3) when did the celebrity tweet? This variation allows us to decompose and measure the relative importance of endorsement, authorship, and credible sourcing effects. Moreover, by varying which celebrities tweet when, we can go further to ask whether an online public health campaign has meaningful offline effects overall.

We study the effects of this induced variation in two ways. First, we use online reactions, i.e., likes and retweets, so we can observe the online reactions of every follower to each tweet.[1] Second, we also study the offline effects of exposure to the campaign by conducting phone surveys of Twitter users. By randomly allocating celebrity activity into one of several phases – either before or after our phone survey – we can examine whether individuals who follow celebrities randomized to tweet before the survey are more likely to have heard of the campaign, updated their beliefs, discussed immunization status with their friends and neighbors, and observed changes in immunization behavior among their friends, relatives, and neighbors.

We chose this setting for several reasons. Twitter is one of the most important mediums of information exchange in the world, with over 1 billion users. Indonesia is quite active on social media; for example, in 2012 its capital, Jakarta, originated the most tweets of any city worldwide. Twitter also has many useful features for our study. Because both the network (i.e., who sees whose tweets) and virtually all information flows over the network (i.e., tweets and retweets) are public, we can precisely map who sees what information, as well as where they saw it, allowing us to observe how much exposure each user had to precise bits of information. By conducting an experiment, randomly varying who tweets what

[1] Note that on Twitter, a "like" corresponds to clicking indicating that one likes the message (which is not pushed to one's followers), while "retweet" subsequently passes on the tweet to all of one's followers. While sometimes individuals retweet tweets they disagree with, adding commentary or simply ironically, "liking" directly conveys approval.

when, we can solve the identification problem of endogenous speaking behavior, as well as disentangle reach vs. endorsement effects. Immunization was chosen as it was a clear public health message, for which celebrities could rely on the Ministry of Health to provide trusted information. And, because of the public nature of Twitter, we can observe people's responses to information both online (by observing their online "liking" and "retweeting" behavior) and offline (by conducting a phone survey of Twitter users and linking their survey responses to whether they were randomly exposed on Twitter to information prior to the survey.)

We build on several facets of the literature. First, an extensive literature in marketing has studied celebrity endorsements. This literature has focused on celebrity endorsements primarily (though not exclusively) in an advertising context, looking at impacts on outcomes such as stock prices (e.g., Agrawal and Kamakura, 1995), sales (e.g., Elberse and Verleun, 2012; Garthwaite, 2014), and brand evaluations, and studying various aspects of the celebrity's identity (e.g., gender, attractiveness); see Bergkvist and Zhou (2016) for a comprehensive review. Our study is one of the first to study celebrity effects in the online space through a real-world, large-scale field experiment, and the first we know of to study public health messaging. Indeed, the only study of a similar magnitude we know of is a marketing study by Gong, Zhang, Zhao, and Jiang (2017), who experimentally vary tweets in China on Sina Weibo asbout TV programs, randomizing whether these tweets were retweeted by influencers. Our study builds on this by decomposing the value of authorship *per se* as opposed to relaying others' messages and identifying the value of credible sources, i.e. health authorities in a public health context. Second, we build on the recent literature on diffusion of information for public policy (e.g, Katona, Zubcsek, and Sarvary, 2011; Banerjee, Chandrasekhar, Duflo, and Jackson, 2013; Beaman, BenYishay, Magruder, and Mobarak, 2016; Beaman and Dillon, 2018). While this literature has studied the flow of information over social networks, and how network position affects the flow of information, it has typically been silent on what aspects of the message matter. Third, we build on the computer science literature on generating online cascades (e.g., Leskovec et al., 2007; Bakshy et al., 2011). Much of this literature suggests seeding messages with many ordinary citizens, as compared to identifying and targeting any particular influencer, but this literature does not generally provide causal evidence on the role of influencers. Finally, our study is complementary to social media experiments that examines how exposure to differing information – such as extent of governmental funding of non-profits, alternative political ideological information, and novel news topics – affects engagement and discourse on the platform (Bail et al., 2018; King et al., 2017; Jilke et al., 2019). Our work takes the message as fixed and studies the messenger: the celebrity.

2. Methods

2.1. **Setting and Sample.** Our study took place in Indonesia in 2015 and 2016. Indonesia is active on social media, ranking third worldwide with 130 million Facebook accounts[2] in 2020 (about half the population), and ranking eighth with 10.6 million Twitter accounts (about 6.4 percent of the population).[3]

The experiment focused on immunization. At the time, Indonesia was trying to improve immunization as part of its drive towards the Millennium Development Goals, so this was a government priority. A set of 550 tweets was developed in coordination with the Ministry of Health that sought to improve information about immunization. The tweets included information about access (e.g., immunizations are free, available at government clinics, and so on); information about immunization's importance (e.g., immunizations are crucial to combat child diseases); and information designed to combat common myths about immunization (e.g., vaccines are made domestically in Indonesia, rather than imported). For each tweet, we identified a source (either a specific link or an organization's Twitter handle). All tweets were approved by the Ministry of Health, and included a common hashtag, #AyoImunisasi ("Let's Immunize"). Each tweet was written in Indonesian, with two versions—one using formal Indonesian, and one using casual/street Indonesian, to match the written tweeting styles of the participants.

With help from the Indonesian Special Ambassador to the United Nations for Millennium Development Goals, we recruited 37 high-profile Twitter users, whom we denote "celebrities," with a total of 11 million Twitter followers. These "celebrities" come from many backgrounds, including music stars, TV personalities, actors and actresses, motivational speakers, government officials, and public intellectuals. They have a mean of 262,647 Twitter followers each, with several having more than one million. While these celebrities primarily tweet about things pertaining to their reason for fame, they also comment occasionally on public issues, often passing on sources or links, so tweets about immunization would not necessarily have been unusual.[4] We also recruited 9 organizations involved in public advocacy and/or health issues in Indonesia with a mean of 132,300 followers each.

In addition, we recruited 1,032 ordinary citizens, primarily Indonesian university students, whom we call "ordinary Joes and Janes". Their role will be to allow us to have essentially

[2] https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/

[3] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[4] For example, three celebrities in our sample (a musician, a TV personality, and a well-known musician's agent) had recently tweeted about the importance of breakfast, including a link to an article about the health benefits of children's breakfast; an athlete tweeted about supporting education for young children; and a musician tweeted in support of Asia Against AIDS.

unimportant, everyday individuals compose tweets that are then retweeted by celebrities. Their Twitter profiles are far more typical, with a mean of 511 followers.

Every participant (celebrities and Joes/Janes) consented to signing up with our app that (1) lets us tweet content from their account (13, 23, or 33 times), (2) randomizes the content of the tweets from a large list of 550 immunization tweets approved by the Ministry of Health, and (3) has no scope for editing.[5] Participants were given two choices: (1) the maximum number of tweets (13, 23, or 33), and (2) a choice of formal or slang language (to better approximate their normal writing style).

2.2. **Experimental Design.** Our experiment is designed to understand which aspects of social media campaigns are important for disseminating a message. The choices we have at our disposal are (a) the originator of the message (a Joe/Jane or a celebrity), (b) whether the message contains a credible source from a public health authority, and (c) the content of the message. Ex ante it may seem obvious, for instance, that sources are better (after all, the information is more credible) and celebrity involvement is better (after all, for a variety of reasons the information may be viewed as more credible). But thinking carefully about the information sharing process demonstrates that, in fact, the effect of each of these design options is actually theoretically ambiguous.[6]

We focus on two main interventions: (1) whether a tweet was tweeted directly by a celebrity, or tweeted by a Joe/Jane and then retweeted by a celebrity; and, (2) for a subset of tweets, whether the tweet included a credible source (i.e., the source link or referring organization's Twitter handle).[7] To ensure everything else was balanced, we also randomized the timing of tweets (which day and what time of day, matching the empirical frequency of local time-of-day of Indonesian tweets), and the content (i.e., which tweet from our pre-prepared bank of approved tweets was tweeted by whom and when).[8] These various randomizations allow us to identify the role of celebrity reach vs. endorsement, as well as the role of celebrity authorship and sourcing.

Randomizing whether a tweet was directly tweeted by a celebrity or retweeted – in combination with the particular way Twitter messages are shown – allows us to decompose the celebrity effects in two ways. First, we use the fact that Twitter messages show the identity

[5] Celebrities could veto a tweet if they did not want it sent from their account, though this in fact never happened.

[6] SI Appendix A presents an application of a simple model developed independently by Chandrasekhar, Golub, and Yang (2018) to demonstrate the ambiguity, though certainly other models can be used and this is inessential for the empirical analysis.

[7] A small subset of tweets on topics deemed 'sensitive' by the Government always included a source; these are excluded from the analysis of sourcing.

[8] Note that in the period we study, a Twitter user saw all tweets and retweets from the users they follow in strict reverse chronological order (i.e., newest tweets appeared first, and so on). Twitter subsequently (in March 2016) applied an algorithm to prioritize the ordering of the tweets, but since in the period we study (July 2015 through February 2016) tweets appeared in strictly chronological order, nothing in our experimental design affects the ordering of tweets in a user's Twitter feed.

of only two people: the originator who wrote the tweet, and the person whom you follow who directly passed it to you (call this person F). Other steps along the chain are omitted. Consider, for example, two chains. One begins with a celebrity (C) tweeting: $C \to F_1 \to F_2$. The other begins with an ordinary Jane/Joe, J, tweeting and the celebrity re-tweeting J's message, i.e.: $J \to C \to F_1 \to F_2$. In the the former case, F_2 sees the celebrity's involvement, whereas in the latter case, F_2 does not. This difference is illustrated in Figure 1. We therefore can study the retweet behavior of F_2's – i.e., followers-of-followers-of-celebrities – across these two cases to identify the endorsement effect premium, i.e. the additional effect of their knowing that C originated the message, holding F_2's network position fixed.[9]

Second, we can decompose the endorsement effect to understand the impact of celebrities speaking in their own voice (an 'authorship' effect). We use the same experimental variation, but look at behavior of the direct followers of celebrities (F_1s), who see both the celebrities' directly-authored tweets and the celebrities' retweets. The second experimental variation, whether a source was retweeted, allows us to identify source effects directly by following the behavior of the ($F_1 s$).

Finally, to measure offline effects, celebrities were randomized into two phases, with half tweeting in the first phase (Phase I, July and August 2015) and half in the second phase (November 2015 - February 2016, Phases II and III). In addition, towards the end of the last phase, all tweets / retweets by a celebrity were then retweeted by a randomly selected number of Joes/Janes (Phase III). We conducted a survey between phases (August - October 2015) of a subset of followers of our celebrities and we use this between-celebrity randomization to estimate the impact of the Twitter campaign on offline beliefs and behavior.

2.3. **Data.** We collected detailed data via the Twitter Firehose and API. Before the experiment began, in early 2015, we collected a baseline image of the publicly available Twitter network, including the list of followers of any celebrity participating in our study.

On Twitter, followers can take two primary actions: "likes" and "retweets". Note that retweets do not necessarily imply endorsement of the views of a tweet, whereas likes do. Second, while likes are public (a user can look up which tweets another user has liked, and can look up who has liked a given tweet), likes are not automatically pushed out as tweets to a user's followers.

For each the of the 672 total tweets originated by our experiment, we tracked each time the tweet was liked or retweeted by any of the over 7.8 million unique users who followed at least one of the participants in our study. When the tweet was retweeted by a celebrity's follower,

[9] A challenge is that the F_1 decision to retweet may be endogenous. We discuss this issue in detail in Section 3.1.1, and show that the results are largely similar in the subset of cases where F_1s were also study participants whom we randomly selected and had retweet exogenously, and hence the sample of exposed F_2s is identical.

we also scraped all of this follower's followers and their liking and retweeting behavior.[10] We denote those retweets / likes coming from a direct follower of a celebrity as F_1 events, and those retweets / likes coming from a follower of a follower of a celebrity as F_2 events. We use the distinction between F_1 and F_2 events in more detail in the analysis below. SI Appendix Table B.1 reports descriptive statistics.

To measure whether online conversations led to offline behavioral changes, we conducted a phone survey on a sample of 2,441 subjects, all of whom followed at least one of our study participants on Twitter. These subjects were recruited primarily via ads placed on the Twitter platform.

2.4. **Estimation.** We estimate three models. First, to estimate the overall effect of endorsement, we focus on the behavior of followers-of-followers-of-celebrities – i.e., F_2's – and estimate by Poisson regression, the equation

$$\mathrm{E}[y_{trcmp}|\mathbf{x}_{trcmp}] = \exp\left(\alpha \cdot \mathrm{Celeb}_{tcm} + \beta \cdot \log(\mathrm{Followers})_r + \omega_c + \omega_m\right) \tag{2.1}$$

where t indexes tweets, r indexes retweeters (i.e., an F_1 who retweeted the tweet t), c indexes celebrities, m indexes the type of message content, and p indexes phase. The variable Celeb_{tcm} is a dummy that takes 1 if the celebrity authored the tweet herself (and hence her identity is visible to F_2), and 0 otherwise (and hence her identity is not visible to F_2). Each observation is a retweet of one of our original tweets, and the dependent variable y_{trcmp} is a count of how many times this retweet was itself either liked or retweeted again by an F_2. Since y is a count, we estimate a Poisson regression, with robust standard errors clustered at the original tweet (t) level. We control for the log number of followers of F_1, and for dummies (ω_m) for different types of messages (e.g., dummies for it being about a fact, importance of immunization, etc). All regressions include celebrity fixed effects (ω_c), which absorb variation in casual/formal style, etc. The coefficient of interest is α, which measures the differential impact of the tweet having been written by the celebrity (as compared to being written by a Joe/Jane) and this being observable to the F_2 deciding whether to retweet.

Second, to decompose the celebrity effect further, we restrict attention to direct followers of the celebrity (F_1 individuals), and estimate

$$\mathrm{E}[y_{tcmp}|\mathbf{x}_{tcmp}] = \exp\left(\alpha \cdot \mathrm{Celeb}_{tcm} + \omega_c + \omega_m\right). \tag{2.2}$$

We now have one observation per tweet, and look at the number of retweets/likes, retweets, or likes by F_1s who are distance 1 from the celebrity. We continue to include celebrity (ω_c), and message-type (ω_m) fixed effects. We run an analogus regression replacing *Celeb* with *Source* to study the impact on F_1 behavior of including public health authority sources.

[10] Since a given user can follow multiple celebrities, the 11 million total followers of celebrities in our sample represents 7.8 million unique users.

Finally, to examine offline effects, we turn to our phone survey data. We define Exposure to Tweets$_i$ as the number of campaign tweets that i is randomized to see through Phase I (normalized to have standard deviation 1). Potential Exposure$_i$ is the total number of campaign tweets that i could potentially see through the campaign given the celebrities she follows at baseline. The experimental design of randomizing celebrities into phases means that, while individuals i may differ in the number of our celebrities they follow, Exposure to Tweets$_i$ is random conditional on Potential Exposure$_i$.

We therefore run logistic regressions of the form

$$f(y_i) = \alpha + \beta \cdot \text{Exposure to Tweets}_i + \gamma \cdot \text{Potential Exposure}_i + \delta' X_i, \tag{2.3}$$

where y_i is the outcome for respondent i and $f(\cdot)$ is log-odds, i.e., $\log\left(\frac{\mathrm{P}(y_i=1|\mathbf{x}_i)}{1-\mathrm{P}(y_i=1|\mathbf{x}_i)}\right)$. X are controls, such as the number of celebrities followed by i, the log of the number of followers of celebrities by i, survey dates, and (in some specifications) demographics and baseline beliefs, selected here and in subsequent regressions by double post-LASSO (Belloni, Chernozhukov, and Hansen, 2014a,b). We report standard errors and p-values clustered at the level of the combination of celebrities followed; further, because of the complex nature of the potential correlation in Exposure to Tweets$_i$ across individuals i induced by partial overlap in which celebrities our survey respondents follow, we present randomization-inference (RI) p-values as well.

3. Results

3.1. **Decomposing The Endorsement Effect: Authorship versus Relaying Messages, and Citing Public Health Authorities.**

3.1.1. *Measuring the Total Endorsement Effect.* We begin by measuring the size of the endorsement effect, and then decompose it into the value that comes from direct authorship of a message as opposed to a more passive involvement, and examine the value of citing public health officials.

To do so, we begin by examining the behavior of F_2s, i.e., followers-of-followers. Recall from the discussion above and the illustration in Figure 1 that if the celebrity retweets a message by a Joe/Jane, and this is retweeted by F_1, F_2 sees the message, sees that it is composed by a Joe/Jane, and knows that F_1 retweeted it. But crucially F_2 does not know that the celebrity had retweeted it: F_2 is likely to be blind to the celebrity's involvement. On the other hand, if the celebrity had written this tweet herself, this would be visible to F_2.

We analyze this by estimating equation (2.1). Table 1 presents the results. We have three main outcome variables: (1) whether the agent either liked or retweeted the tweet, (2) whether an agent liked the tweet, and (3) whether an agent retweeted the tweet. Columns 1, 3, and 5 present the results on the full sample for each of these dependent variables.

We see large endorsement effects. Having a celebrity compose and tweet the message relative to having a Joe/Jane compose the message and the celebrity retweet it leads to a 72 percent (0.54 log point) increase in the retweet or like rate (column 1, $p = 0.001$; note that since this is a Poisson model, the coefficients are interpretable as the change in log number of retweets/likes) by followers-of-followers (F_2s). The results are qualitatively similar for retweets and likes alone—68 percent (0.52 log points) for retweets and 92 percent (0.66 log points) for likes.

These results imply that, holding the content of the tweet constant (since it is randomized across tweets) and holding the F_2 position in the network constant (since they are all followers-of-followers of the celebrity), having the F_2 be aware of the celebrity's involvement in passing on the message substantially increases the likelihood that the F_2 responds online.

The results allow us to begin to decompose the reason for retweeting. Specifically, the estimates imply that 63-72 percent[11] of retweets come from the fact that the celebrity is involved (in this case having written the tweet), with the remainder coming from the intrinsic interest in the content of the tweet itself.

We document similar effects of an organization being the originator rather than a Joe/Jane in Table E.1 of Appendix E.[12]

There are two main potential threats to identification. First, when we look at F_2 agents, i.e., those who are at distance two from the celebrity, whether a given agent sees a retweet from his or her F_1s may be endogenous and respond to our treatment, i.e., which F_1s choose to retweet the message may be directly affected by the fact that the celebrity composed the message. In equation (2.1), we control for the log number of followers of the F_1 who retweeted the message, and hence the number of F_2s who could potentially retweet it, so there is no mechanical reason for a bias in equation (2.1). But there may nevertheless be a *compositional* difference in which F_1s retweet it, which could potentially lead to selection bias of which F_2s are more likely to see the retweet.

To address this issue, in the last phase of the experiment, we added an additional randomization. We use the subset of Joes/Janes who are also F_1s, and so direct followers of our celebrities. For some of these Joes/Janes, we randomly had their accounts retweet our celebrities' tweets and retweets in the experiment; that is, we created exogenous F_1s. For this sample, we can look at how *their* followers – that is, the followers of F_1 Joes/Janes we exogenously forced to retweet a particular tweet – responded as we randomly vary whether

[11] $\frac{\exp(\alpha)}{1+\exp(\alpha)}$ for coefficient α.

[12] Recall we only have 9 organizations, which reduces the overall instances of such cases, so we relegate this to an appendix. Also, we condition on non-sensitive tweets for this sample.

the celebrity, an organization, or a Joe/Jane composes the message. We analyze this experiment by estimating equation (2.1) just as we did for the full sample, but here we know that whether an F_2 sees the tweet is exogenous by construction.

Columns 2, 4, and 6 present the results. The point estimates are if anything somewhat larger than in the full sample, and we cannot reject equality. Statistical significance is reduced somewhat (p-values of 0.119, 0.111, and 0.107 in columns 2, 4, and 6 respectively), but the fact that results are broadly similar to the overall effects in columns 1, 3, and 5 suggests that the possible endogenous selection of F_1s in our full sample is not leading to substantial bias.

The second potential confound comes from the fact that a retweet shows how many times the original tweet has been retweeted or liked when the user views it (see Figure F.1; the number of retweets is next to the arrow, and the number of likes is next to the heart). Since our treatment assignment affects the retweet count, this itself could spur further changes in the likelihood of retweeting. The same randomization of forced Joe/Jane retweets also helps address this issue, because we randomly varied the number of Joes/Janes we forced to retweet a particular tweet. We find that being randomly assigned one, five, ten, or even fifteen extra retweets makes no impact on the number of F_1 or F_2 retweets that the given tweet faces (see Appendix D, Table D.1).

3.1.2. *Decomposing Endorsement: Authorship vs. Passing on Messages of Others.* The preceding analysis compared individuals who were effectively randomly blinded to whether a celebrity was or was not involved in the message composition and message passing to estimate an endorsement effect. While celebrity involvement was through authorship in that case, we do not know to what extent *involvement* per se or *authorship* per se mattered.

To decompose this, we can examine the behavior of the direct followers of the celebrities themselves (i.e., F_1s). For F_1s, they know the celebrity is involved either way, but the randomization changes whether it was directly authored by the celebrity or retweeted.

We estimate equation 2.2, and present results in Panel A of Table 2. We find that authorship is important: tweets authored by celebrities are 200 percent more likely to be retweeted/liked than those where the celebrity retweets (column 1, $p < 0.001$). An agent who observes a tweet composed by the celebrity rather than a retweet of a Joe/Jane is 120 percent more likely to like it (column 2, $p < 0.001$) and 280 percent more likely to retweet it (column 3, $p < 0.001$).

This fact allows us to further decompose the impacts of celebrity. These estimates imply that 79 percent of the endorsement effect estimated earlier comes from authorship per se. Combining these estimates with those in the previous section suggests that, on net, 56 percent of the celebrity effect comes from authorship, 14 percent from endorsement, with the remainder attributable to the intrinsic interest of the message.

3.1.3. *Citing Public Health Authorities.* Finally, we examine the impact of citing sources. Every tweet in our databank was paired with a source, but we randomized at the tweet level whether this source was included or not in the tweet. These source citations in our context come in several forms. In some cases, the tweet refers to the website or Twitter handle of a trusted authority who has issued that statement. For example, one tweet says "Polio vaccine should be given 4 times at months 1, 2, 3, 4. Are your baby's polio vaccines complete? @puskomdepkes' where "@puskomdepkes" is a link to the Twitter handle of the Ministry of Health (known as *DepKes* in Indonesian). In other cases, explicit sources are cited, with a Google shortened link provided.[13]

We re-estimate equation (2.2) at the F_1 level, adding a variable for whether the tweet was randomly selected to include a source.[14] Panel B of Table 2 presents the results.

On average, we find that citing a public health authority reduces the retweet and liking rate by 26.3 percent (-0.306 log points; $p = 0.051$, column 1). We find that both the retweeting rate and liking rate experience broadly similar declines (retweeting declines by 0.318 log points, $p = 0.048$, column 2; liking declines by 0.277 log points, $p = 0.13$, column 3, not significantly different from zero). On net, the results show that when a message is relayed by citing a public health authority, willingness to pass it on downstream, and perhaps liking the message, declines.

3.2. **Does Online Discussion Have Offline Effects?**

3.2.1. *Did people hear about the campaign?* We next examine whether an online celebrity endorsement campaign can begin to have measurable offline effects. To investigate this, we estimate equation (2.3). We use the fact that our offline survey was conducted between Phases I and II, as described in Section 2, so that conditional on the number of our celebrities a user followed, exposure to our campaign as of the time of the phone survey was randomly assigned.

We begin with what can be thought of as akin to a first-stage in Panel A of Table 3. We ask whether respondents were more likely to have heard of our hashtag (#AyoImunisasi) or heard about immunization discussions from Twitter if they were randomly more exposed to campaign tweets, conditional on their potential exposure.

[13] Note that Twitter automatically produces a short preview of the content if the site linked to has Twitter cards set up. There is one non-Google shortened link used when citing IDAI (Ikatan Dokter Anak Indonesia, the Indonesian Pediatric Society).

[14] Note that the number of observations is smaller, because some tweets on topics deemed 'sensitive' by the Government always included a source, as noted above. We restrict analysis to tweets for which we randomized whether the source was included.

Column 1 examines whether the respondent heard of #AyoImunisasi (and therefore remembers the exact hashtag) and column 2 examines whether the respondent heard of a discussion of immunization on Twitter.

We find that a one-standard deviation increase in exposure to the campaign (15 tweets) corresponds to a 16.75 percent increase in the probability that the respondent had heard of our hashtag relative to a mean of 7.7 percent (clustered $p = 0.044$, RI $p = 0.107$).[15] Further, a one-standard deviation increase in exposure corresponds to a 8.3 percent increase in the probability they heard about immunization in general from Twitter relative to a mean of 18.1 percent (clustered $p = 0.106$, RI $p = 0.046$).

3.2.2. *Did people then increase their knowledge about immunization facts?* Next we ask whether exposure to the campaign led to increased knowledge about immunization. Our survey asks questions about several categories of knowledge. First, we check for knowledge of several common "myths" about vaccination that our campaign tried to cover. In particular, we ask whether people know that vaccines are domestically produced, to combat the common rumor in Indonesia that they contain pig products in production (which would make them unacceptable for Muslims, who represent the vast majority of Indonesia's population; domestic products are known to be halal). Second, we ask whether they believe that natural alternatives (breastfeeding, herbal supplements, alternative supplements) replace the need for immunization. Third, we ask whether they are aware that typical symptoms (mild fevers or swelling) are to be expected and not a cause for alarm. The second category we ask about is "access" information; in particular, we ask whether they know that it is free to get one's child vaccinated at government health centers. All of these issues were covered in the campaign.

The fact that we emphasize the role of messages that dispel myths, such as the fact that vaccines are not domestically produced, is echoed in online behavior. In particular, Table C.1 in Appendix C shows that tweets concerning myths diffused more widely than facts that were non-myths, meaning that the exposure would have been more about myth-busting facts. Moreover, myth-dispelling facts comprised 36.7 percent of all tweets and 82.4 percent of all fact-related tweets sent out (i.e., myths compared to other facts).

Table 3, Panel B, presents the results for each of these four categories of information. We find knowledge effects for our most prominent tweet—domestic production (column 1)—though not on rumors about substitutability, side-effects, nor free access (columns 2-4). Seeing 15 campaign tweets in general corresponded to an increase of 5 percent in the probability of correctly answering the domestic question on a base of 57.6 percent (clustered $p = 0.042$,

[15] Note that the table reports impacts on log-odds; we report marginal effects in the text, which are interpretable as percent increases. For example, the 16.75 percent increase in column 1 corresponds to an increase in 0.197 in log-odds.

RI $p = 0.028$; if we adjust for the fact that 4 questions are asked using a Bonferroni-style adjustment; these p-values would be 0.168 and 0.112, respectively).

3.2.3. *Communication about immunization.* We begin by asking whether individuals were more likely to know about what their neighbors, friends, and relatives' immunization behavior was, which would be a byproduct of offline conversations since June 2015, to capture the campaign's effect. Immunizations in Indonesia take place at monthly *posyandu* meetings, which occur each month in each neighborhood (usually hamlets, or *dusun*, in rural areas, and neighborhoods known as *rukun warga*, or *RW*, in urban areas; see Olken, Onishi, and Wong (2014)), so if knowledge would increase, one might expect it to be the knowledge about immunization practices of ones' neighbors.[16]

Panel A of Table 4 presents the results. Column 1 shows that being exposed to 15 more tweets corresponds to a 5.2 percent increase in the probability of knowing the neighbor's status (clustered $p = 0.004$, RI $p = 0.088$). We do not consistently see significant effects on non-neighbor friends. With relatives the point estimates are comparably large, though the estimates are noisier. In any case, this is also consistent with the idea that for the most part, individuals have a higher rate of knowledge of health status of those they have closer relationships with (e.g., relatives) and that this is unlikely to change.

3.2.4. *Did exposure lead to changes in immunization rate?* Finally, we ask whether, among those who knew their friends, neighbors, or relatives' immunization behavior, is there more immunizing behavior since June 2015, when there is more exposure to the campaign? That is, does the campaign appear to change immunization behavior as reported by our survey respondents?

Panel B of Table 4 presents the results, where the dependent variable is whether the respondent knows of anyone among their friends, neighbors, or family who immunized a child. We condition the sample when we look at knowledge of immunization among neighbors, friends, and family to those who knew whether the vaccination status of the children of members of these individuals, so this effect is distinct from the effects reported above. Then we ask whether exposure to treatment is associated with greater incidence of vaccination take-up among neighbors, friends, and family.

An important caveat to this table is that knowledge of neighbors', friends', and family members' children's vaccination status itself is affected by treatment (as discussed above). This is particularly true for neighbors. If knowledge is itself affected by treatment, there can be at least two natural interpretations of these results. The first interpretation is that if the increase in reporting one's child's vaccination status to one's friend is unrelated to the actual status itself, then this regression picks up the treatment effect on vaccination

[16] Note that here the sample is restricted to respondents who know friends, relatives, or neighbors with at least one child (ages 0-5) respectively as this is the relevant set, which reduces the survey sample.

itself. The second interpretation is that the treatment causes differential reporting of status to one's neighbors, so those who took up vaccination for their children may be more likely to speak about it. We cannot separate between the two interpretations, but we present the results nonetheless, advising that the estimates are to be interpreted with some caution.

Column 1 shows that when looking at neighbors, an increased exposure by 15 tweets corresponds to a 12.5 percent increase in the number of vaccinations (clustered $p = 0.071$, RI $p = 0.132$) relative to a mean of 0.356. For friends, shown in column 2, we find an increased exposure of 15 tweets corresponds to a 16.0 percent increase in the number of vaccinations (clustered $p = 0.001$, RI $p = 0.071$) relative to a mean of 0.353. Column 3 presents results looking at relatives. An increased exposure by 15 tweets corresponds to a 9.6 percent increase in the number of vaccinations among relatives (clustered $p = 0.159$, RI $p = 0.06$) relative to a mean of 0.314. Finally, Column 4 looks at own behavior, and our estimate is not statistically different from zero. On net, the results in this section are suggestive, but not dispositive, that an online Twitter campaign can have offline effects on knowledge and, potentially, behavior.

4. Discussion

Our results allow us to study how to design a public health campaign to effectively deploy celebrities by conducting a large-scale, online field experiment. We are able to decompose to what extent celebrity authorship *per se* as opposed to passing on messages of others matters, and to show in particular whether citing public health officials amplifies or reduces these effects. We find the vast majority (79%) of the 70% increase in retweet rate due to celebrity endorsement coming from authorship.

We also find that explicitly referring to public health sources has an adverse effect. This result may seem surprising, since one might expect that a sourced message may be more reliable. There are, however, several possible explanations for this finding. One idea, which is certainly not the only one, is that for an F_1, passing on a message has both instrumental value (delivering a good message), as well as a signaling value (conveying to followers that the F_1 is able to discern which information is good). This model was developed by the authors in prior work (Chandrasekhar et al., 2018). We develop this model in this context in SI Appendix A. Other stories are certainly possible as well. At a broader level, these the results are consistent with the results on celebrity authorship: messages are most likely to be passed on when they come from the celebrity speaking directly, rather than passing on messages from others.

We also find offline effects: messages are heard, certain beliefs change while others do not, and offline communication about vaccination increases as does reported immunization behavior among respondents' neighbors, friends, and relatives. Specifically, being randomly exposed to only 15 more campaign tweets leads to 10-20% increased awareness. Further, knowledge of the domestic production of vaccines increases – a key anti-myth message in this context – increases, consistent with anti-myth messages diffusing more than other messages online. At a minimum this exposure leads to increased communication about vaccination among one's neighbors, friends, and relatives leading to greater knowledge of their decisions, with suggestive results suggesting increases in immunization behavior as well.

These results may be useful in designing information campaigns in conjunction with the COVID-19 crisis, such as to encourage social distancing or, later, to encourage immunization if and when a vaccine becomes available. In so doing, the key message of our findings is that an effective design includes recruiting influential agents, like celebrities, to send self-authored messages in their own voice, *without* explicitly citing credible public health sources.

References

Agrawal, J. and W. A. Kamakura (1995): "The economic worth of celebrity endorsers: An event study analysis," *Journal of marketing*, 59, 56–62. 1

Bail, C. A., L. P. Argyle, T. W. Brown, J. P. Bumpus, H. Chen, M. F. Hunzaker, J. Lee, M. Mann, F. Merhout, and A. Volfovsky (2018): "Exposure to opposing views on social media can increase political polarization," *Proceedings of the National Academy of Sciences*, 115, 9216–9221. 1

Bakshy, E., J. M. Hofman, W. A. Mason, and D. J. Watts (2011): "Everyone's an influencer: quantifying influence on twitter," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, 65–74. 1

Banerjee, A., A. G. Chandrasekhar, E. Duflo, and M. O. Jackson (2013): "The diffusion of microfinance," *Science*, 341, 1236498. 1

Banerjee, A. V., E. Breza, A. G. Chandrasekhar, and B. Golub (2018): "When Less is More: Experimental Evidence on Information Delivery During India's Demonetization," . A.1

Beaman, L., A. BenYishay, J. Magruder, and A. M. Mobarak (2016): "Can Network Theory based Targeting Increase Technology Adoption?" *Working Paper*. 1

Beaman, L. and A. Dillon (2018): "Diffusion of agricultural information within social networks: Evidence on gender inequalities from Mali," *Journal of Development Economics*, 133, 147–161. 1

Belloni, A., V. Chernozhukov, and C. Hansen (2014a): "High-dimensional methods and inference on structural and treatment effects," *The Journal of Economic Perspectives*, 29–50. 2.4

——— (2014b): "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81, 608–650. 2.4

Bergkvist, L. and K. Q. Zhou (2016): "Celebrity endorsements: a literature review and research agenda," *International Journal of Advertising*, 35, 642–663. 1

Bursztyn, L., G. Egorov, and R. Jensen (2017): "Cool to be smart or smart to be cool? Understanding peer pressure in education," *The Review of Economic Studies*. A.1

Bursztyn, L. and R. Jensen (2015): "How does peer pressure affect educational investments?" *Quarterly Journal of Economics*, 130, 1329–1367. A.1

Chandrasekhar, A. G., B. Golub, and H. Yang (2018): "Signaling, Shame, and Silence in Social Learning," Tech. rep., National Bureau of Economic Research. 6, 4, A.1, A.3

Elberse, A. and J. Verleun (2012): "The economic value of celebrity endorsements," *Journal of advertising Research*, 52, 149–165. 1

GARTHWAITE, C. L. (2014): "Demand spillovers, combative advertising, and celebrity endorsements," *American Economic Journal: Applied Economics*, 6, 76–104. 1

GONG, S., J. ZHANG, P. ZHAO, AND X. JIANG (2017): "Tweeting as a marketing tool: A field experiment in the TV industry," *Journal of Marketing Research*, 54, 833–850. 1

JILKE, S., J. LU, C. XU, AND S. SHINOHARA (2019): "Using large-scale social media experiments in public administration: Assessing charitable consequences of government funding of nonprofits," *Journal of Public Administration Research and Theory*, 29, 627–639. 1

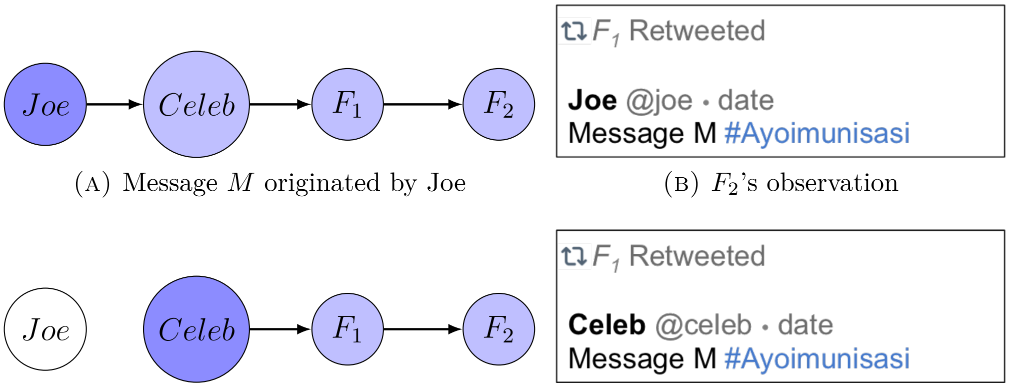
KATONA, Z., P. P. ZUBCSEK, AND M. SARVARY (2011): "Network Effects and Personal Influences: The Diffusion of an Online Social Network," *Journal of Marketing Research*, 48:3, 425–443. 1

KING, G., B. SCHNEER, AND A. WHITE (2017): "How the news media activate public expression and influence national agendas," *Science*, 358, 776–780. 1

LESKOVEC, J., L. A. ADAMIC, AND B. A. HUBERMAN (2007): "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, 1, 5. 1

OLKEN, B. A., J. ONISHI, AND S. WONG (2014): "Should Aid Reward Performance? Evidence from a field experiment on health and education in Indonesia," *American Economic Journal: Applied Economics*, 6, 1–34. 3.2.3

Figures



F_1 Retweeted

Joe @joe • date
Message M #Ayoimunisasi

(a) Message M originated by Joe

(b) F_2's observation

F_1 Retweeted

Celeb @celeb • date
Message M #Ayoimunisasi

(c) Message M originated by Celeb

(d) F_2's observation

Figure 1. Identification of the value of endorsement of celebrity involvement.

Tables

Table 1. Estimating the average value of celebrity involvement using followers-of-followers' behavior

	(1)	(2)	(3)	(4)	(5)	(6)
	Poisson	Poisson	Poisson	Poisson	Poisson	Poisson
VARIABLES	# Pooled	# Pooled	# Retweets	# Retweets	# Likes	# Likes
Celeb writes and tweets	0.544	0.788	0.518	0.931	0.664	1.109
	(0.166)	(0.505)	(0.166)	(0.584)	(0.482)	(0.687)
	[0.00105]	[0.119]	[0.00175]	[0.111]	[0.168]	[0.107]
Observations	1,997	911	1,997	911	1,997	911
Joe/Jane writes mean	0.0417	0.00915	0.0417	0.00686	0.00745	0.00229
Forced Joes/Janes only		✓		✓		✓

Notes: Standard errors (clustered at the original tweet level) are reported in parentheses. *p*-values are reported in brackets. Sample conditions on all tweets originated by Joes/Janes or celebrities. All regressions control for phase, celebrity fixed effects, content fixed effects, and the log number of followers of the F_1.

Table 2. Estimating the value of authorship and citing public health officials using followers' behavior

Panel A: Decomposing what % of the involvement effect comes from authorship

	(1)	(2)	(3)
	Poisson	Poisson	Poisson
VARIABLES	# Pooled	# Retweets	# Likes
Celeb writes and tweets	1.101	1.329	0.803
	(0.0840)	(0.0910)	(0.105)
	[0]	[0]	[0]
Observations	451	451	451
Joe/Jane writes and Celeb retweets mean	2.058	1.045	1.013

Panel B: Measuring the effect of citing public health sources

	(1)	(2)	(3)
	Poisson	Poisson	Poisson
VARIABLES	# Pooled	# Retweets	# Likes
Source cited	-0.306	-0.318	-0.277
	(0.157)	(0.161)	(0.183)
	[0.0513]	[0.0478]	[0.130]
Observations	492	492	492
Depvar Mean	3.644	3.644	3.644

Notes: In Panel A, sample conditions on all tweets originated by Joes/Janes or celebrities. In Panel B, the sample conditions ono non-sensitive tweets. All regressions control for phase, celebrity fixed effects, content fixed effects. Standard errors (clustered at the celebrity/organization level) are reported in parentheses.

TABLE 3. Knowledge of Campaign and Facts

Panel A: Did people offline hear about the campaign?

	(1)	(2)
	Logit	Logit
	Heard of	Heard of
VARIABLES	#*Ayoimunisasi*	immunization from Twitter
Std. Exposure to tweets	0.197	0.108
	(0.0980)	(0.0666)
	[0.0443]	[0.106]
	{.107}	{.046}
Observations	2,164	2,404
Potential exposure control	✓	✓
Double Post-LASSO	✓	✓
Depvar Mean	0.0776	0.181

Panel B: Did people offline increase knowledge?

	(1)	(2)	(3)	(4)
	Logit	Logit	Logit	Logit
VARIABLES	Domestic	Substitutes	Side-effects	Free
Std. Exposure to tweets	0.120	-0.0391	0.0305	0.0549
	(0.0591)	(0.0589)	(0.0624)	(0.0687)
	[0.0424]	[0.506]	[0.625]	[0.424]
	{.028}	{.891}	{.751}	{.629}
Observations	2,434	2,440	2,440	2,440
Potential exposure control	✓	✓	✓	✓
Double Post-LASSO	✓	✓	✓	✓
Depvar Mean	0.576	0.527	0.486	0.680

Notes: Standard errors (clustered at the combination of celebs followed level) are reported in parentheses. Clustered p-values are reported in brackets. Randomization inference (RI) p-values are reported in braces. Demographic controls include age, sex, province, dummy for urban area and dummy for having children. One standard deviation of exposure is 14.96 tweets.

Table 4. Communication With and Behavior of Neighbors, Friends, and Relatives

Panel A: Communication: Knowledge of immunization behavior of others

	(1)	(2)	(3)
	Logit	Logit	Logit
VARIABLES	Neighbor	Friend	Relative
Std. Exposure to tweets	0.231	0.0156	0.214
	(0.0814)	(0.0826)	(0.132)
	[0.00449]	[0.850]	[0.105]
	{.088}	{.778}	{.462}
Observations	1,642	1,626	1,564
Potential exposure control	✓	✓	✓
Double Post-LASSO	✓	✓	✓
Depvar Mean	0.775	0.813	0.923

Panel B: Immunization behavior of others and self

	(1)	(2)	(3)	(4)
	Logit	Logit	Logit	Logit
VARIABLES	Neighbor	Friend	Relative	Own
Std. Exposure to tweets	0.194	0.246	0.140	-0.0840
	(0.107)	(0.0955)	(0.0997)	(0.0886)
	[0.0707]	[0.00994]	[0.159]	[0.343]
	{.132}	{.071}	{.06}	{.66}
Observations	682	682	682	634
Potential exposure control	✓	✓	✓	✓
Double post-LASSO	✓	✓	✓	✓
Depvar Mean	0.356	0.353	0.314	0.486

Notes: In both panels, standard errors (clustered at the combination of celebs followed level) are reported in parentheses. Clustered p-values are reported in brackets. Randomization inference (RI) p-values are reported in braces. Demographic controls include age, sex, province, dummy for urban area and dummy for having children. One standard deviation of exposure is 14.96 tweets. In Panel A, the sample is restricted to respondents who know friends/relatives/neighbors with at least one child (ages 0-5) respectively. In Panel B, the sample when looking at network members' behaviors (columns 1-3) is restricted to respondents who know the behavior of their network. When looking at own behavior in column 4, sample restricted to respondents with children younger than age 2.