CELEBRITIES AND PUBLIC HEALTH CAMPAIGNS: A NATIONWIDE TWITTER EXPERIMENT PROMOTING VACCINATION IN INDONESIA

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

Abstract. We ask whether celebrities can help spread information about public health, above and beyond the fact that their statements are seen by many, and ask how they can be most effective in doing so. We conducted a nationwide Twitter experiment with 46 high-profile Indonesian celebrities and organizations, with over 11 million followers, who agreed to randomly tweet or retweet content promoting immunization. Our design exploits the structure of what information is passed on along a retweet chain on Twitter to decompose how celebrities matter. We find that messages that can be identified as being authored by celebrities are 72 percent more likely to be passed or liked compared to similar messages without a celebrity imprimatur. Decomposing this effect, we find that 79 percent of the celebrity effect comes from the act of celebrity authorship itself, as opposed to merely passing on a message. Explicitly citing an external source decreases the likelihood of passing the message by 27 percent. The results suggest that celebrities have an outsize influence in shaping public opinion, particularly when they speak in their own voice.

JEL Classification Codes: D83, I15, O33

Date: March 12, 2020.

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 online followers, celebrities have a direct channel to spread messages on a wide variety of issues, many of which are far removed from their original reason for fame. Their very 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 of all time, with 41 million uses as of September 2018. Prominent celebrity users include LeBron James, Kim Kardashian, Kanye West, and Serena Williams, among many others. In public health, the #IceBucketChallenge, promoting awareness of Lou Gehrig's disease, became the sixth most-used social issue hashtag of all time following participation by many celebrities, from Oprah to Bill Gates. Each of these campaigns was initiated by a less-well-known activist, but was made prominent in part through celebrity participation. Policymakers and firms therefore often seek out celebrity endorsements, whether to advance public-interest causes or to promote products.

The effective design of a public health campaign depends crucially on the details as to why and how celebrities are influential. Celebrities have broad *reach*. Many people are watching what Kim Kardashian says or does, and hence her actions and utterances are seen by many people. However, the fact that a celebrity per se spoke on a topic may have an impact above and beyond that of distributing the same message to the same recipients. We call this additional effect – above and beyond what would have happened had the message been delivered without the celebrity's imprimatur – an *endorsement* effect. That is, the fact that it was Kim Kardashian who was willing to message about a cause or product may lead people to change their behavior. They may pass on the message because they update about the quality of the product or the importance of that cause, or people may simply want to be like her or demonstrate that they participate in new social trends, among many other reasons. If this endorsement effect is present, this means that celebrities have an outsized importance: it is not just that they reach so many people, but their decision to speak *per se* has an additional effect.

Conditional on deciding to engage on a topic, celebrities face several choices, particularly with regard to social issues outside their core domain of expertise. In particular, they can speak in their own voice, they can pass on the message of others, and/or they can seek to bolster their credibility by referring to an external source for the information they are sharing. Each of these choices could bolster or dampen the celebrities' endorsement effect. These decisions matter for policy design, such as in a public health campaign, as they could affect how governments encourage celebrities to share messages in order to maximize their effectiveness in generating the general public's engagement.

Understanding these questions is challenging. First, celebrities' decisions about whether to make public statements, and what they say, are all choice variables and influenced by the general information environment into which they are speaking. People also consume information from such a wide variety of sources that in most contexts it is also nearly impossible to isolate the response to a particular piece of information. Even if one could credibly solve the endogeneity problem of whether and how celebrities choose to speak on a topic, and could isolate the impact on a particular individual, a given action by a celebrity bundles reach and endorsement effects, making it hard to disentangle why, precisely, these 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 each of these tweets was randomly chosen by us from a set of tweets approved by the Indonesian Ministry of Health, all of which featured a campaign hashtag #AyoImunisasi ("Let's Immunize"). All our participants joined knowing that they would not be able to affect the text or timing of the tweets.[1]

The experiment randomly varied the tweets along two key dimensions: (1) Did the celebrity / organization send the tweet from their account, or did they retweet a message (drawn randomly from the same tweet library) sent by us from an ordinary (non-celebrity) user's account?; and (2) Did the tweet explicitly cite a source to bolster its credibility? This random variation allows us understand whether and why celebrity-involved campaigns have influence – differentiating endorsement from reach, and then decomposing the endorsement effect into the effect of speaking in one's own voice versus passing on messages of others. We study the effects of this induced variation using online reactions to the tweets, i.e., likes and retweets, so we can observe the online reactions of every individual follower to every specific tweet.[2]

We chose this setting for several reasons. Indonesia is very active on social media; for example, in 2012 its capital, Jakarta, originated the most tweets of any city in the world. Twitter also has a number of 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) on Twitter are public information, we can precisely map which individual sees

[1]Celebrities were allowed to a veto a tweet if they did not want it sent from their account, though this in fact never happened during the campaign.

[2]Note that on the Twitter platform, a "like" corresponds to simply clicking a button to indicate that one likes the message (and the action is not pushed to one's followers), while "retweet" subsequently passes on the tweet to all of one's followers. While it is certainly the case, *a priori*, that individuals may retweet tweets that they even disagree with, adding commentary or simply ironically, "liking" the tweet directly conveys approval.

what information, as well as where they saw it from, allowing us to observe for each user how much exposure they had to precise bits of information. By conducting an experiment, in which we randomly vary who tweets what when, we can both 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 they could share.

Twitter is also one of the most important mediums of information exchange in the world. With over 1 billion users and 328 million active users, Twitter provides a platform for individuals to broadcast messages widely. Celebrities, politicians, and organizations are widely followed.[3] As such, influencers have a platform to directly message en masse and engage on timely issues.

We begin by using our design, in combination with the unique structure of how information is passed on Twitter, to distinguish reach from endorsement. Messages in Twitter are passed on by retweeting a message to one's followers. Crucially for our design, when a message is retweeted, the follower observes who originally composed the tweet, and who retweeted it directly to the follower, but not any intermediate steps in the path.

To see how this allows us to distinguish reach from endorsement, consider the difference between what happens when 1) we have a celebrity directly compose and tweet a message, compared to 2) when we have a celebrity retweet a message drawn from the same pool of tweets but originated by a normal citizen (whom we henceforth denote as a "ordinary Joes and Janes"; these Joes/Janes are also participants on our campaigns). In the first case, some celebrity followers (whom we denote F_1) retweet it to their followers, whom we denote F_2. The followers-of-followers (F_2s) observe that the celebrity authored the message and that F_1 retweeted it. But in the second case, when the celebrity retweeted a Joe/Jane's message rather than composed it herself, the followers-of-followers of the celebrity (F_2s) observe only that the Joe/Jane tweeted and that F_1 then retweeted for F_2 to see. Notice that in this way, F_2 is randomly blinded to the celebrity's involvement in the latter case, as compared to the former: differences in F_2's behavior therefore correspond to differences due to knowing that the celebrity was involved.[4] In the period we study, the ordering of the Twitter feed was strictly chronological, so this design manipulates whether the F_2s know about the celebrities involvement without affecting how prominently the message appeared in the Twitter feed.

Using this design, we find a substantial endorsement effect. Specifically, when an individual observes a given message through a retweet, and that message was randomized to

[3]Among the most followed worldwide are Barack Obama (114 million), Katy Perry (108 million), Cristiano Ronaldo (83 million), Donald Trump (74 million), and Kim Kardashian (64 million) as of March 2020.

[4]A challenge in the design is that the F_1 decision to retweet may be endogenous. We discuss this issue in detail in Section 3.1 below, 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.

be originally composed by a celebrity as compared to an ordinary individual, there is a 72 percent increase in the number of likes and retweets, compared to similar messages when the celebrity's involvement was masked. We find similar results even when we restrict attention to those cases where $F_1 s$ are participants in the experiment and we exogenously had them retweet the message, ensuring that whether the F_2 was exposed to the message in the first place was completely exogenous. This implies that the the endorsement effect corresponds to 63-72% of the reason as to why a downstream follower (F_2) retweets a celebrity.

The preceding analysis does not distinguish between whether celebrity endorsement matters because a downstream individual knows that a celebrity was involved per se, or because the celebrity actually authored the message. We then seek to decompose the endorsement effect further to understand the impact of celebrity speaking in their own voice (an 'authorship' effect). We use the same experimental variation, but now look at behavior of the direct followers of celebrities (F_1s), who see both the celebrities' directly authored tweets and the celebrities' retweets. Tweets directly authored by celebrities are 280 percent more likely to be retweeted than comparable messages passed on by comparable celebrities but authored by ordinary users. The vast majority – 79 percent – of the endorsement effect therefore comes from celebrities speaking in their own voice.

Celebrities also may choose to bolster their credibility on topics outside their core expertise by explicitly citing sources. We find that this approach actually reduces message diffusion: messages are less likely to be passed in our experiment if they are randomly assigned to include source. This is true regardless of whether the tweet was composed by the celebrity themselves, or composed by an ordinary Joe/Jane and retweeted by a celebrity. The magnitudes are substantial: randomly attaching a source to a tweet reduces the probability of retweeted by 27 percent. One interpretation is the information is less novel if it is sourced; more generally, in Online Appendix A we discuss theoretically how increasing the reliability of information passed has ex ante ambiguous effects on the probability the information is passed.

Taken together, our estimates allows us to decompose the celebrity effect. On net, we estimate that 56 percent of the celebrity effect comes from authorship, 14 percent from the endorsement value, and the remainder is attributable to the intrinsic interest of the message itself. The results suggest that celebrities can play an important role in the diffusion of public health messages, but to do so, they need to speak in their own voice.

Related Literature. This work relates to a literature on the diffusion of information for public policy (e.g, Ryan and Gross, 1943; Kremer and Miguel, 2007; Katona, Zubcsek, and Sarvary, 2011; Banerjee, Chandrasekhar, Duflo, and Jackson, 2013; Beaman, BenYishay, Magruder, and Mobarak, 2016). Our paper represents one of the only large-scale randomized controlled trials of an online diffusion experiment, particularly one that involves major

influencers.[5] Moreover, while this literature has studied the flow of information over social networks, and how position in the network affects the flow of information, it has typically been silent on whether the identity of the individual who passes on the information matters per se.[6] This is because normally the identity of an individual and that individual's position in the network go hand-in-hand, so varying who is sending the information changes both of these simultaneously; our experimental design, by contrast, allows us to separate these two effects. Unlike these previous studies, our study represents exactly the kind of large-scale public awareness campaign that governments and large-scale policymakers are interested in, as represented by the Ministry of Health's and World Bank's interest in partnership. To our knowledge this is the only study looking to decompose the reason a celebrity's messages is passed in: parsing authorship, involvement per se, and choice of referencing supplemental credible information.

There is also a literature on generating online cascades (Leskovec et al., 2007; Bakshy et al., 2011). This literature follows online diffusions through Twitter, Facebook, and other social media, and through observational studies looks at what drives and does not drive diffusion. Much of the literature concludes that under a wide range of assumptions, it is more worthwhile to seed a message with many ordinary citizens as compared to identifying and targeting any particular influencer. As the research notes, however, there is no causal evidence for the role of influencers here, and certainly no causal evidence to parse what aspects of celebrity involvement matters. Of course, in observational studies, what celebrities say, whether they cite sources, and whether they endorse others' messages are all endogenous. Our experimental design allows us to answer these questions.

Finally, our study is complementary to others that could inform the design of a social media campaign. These experiments on social media (Facebook and Twitter) look at how exposure to differing information—such as extent of governmental funding of non-profits, alternative political ideological information, 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. We alter whether a celebrity simply passes on a message or actually authors it, and also whether the celebrity appeals to external credible sources.

[5]The only study of a similar magnitude we are aware of is Gong, Zhang, Zhao, and Jiang (2017), who experimentally vary tweets in China on Sina Weibo about TV programs, and randomized whether these tweets were retweeted by influencers with large numbers of followers. That study does not seek to unpack reach vs. endorsement effects or the value of sources, however.

[6]An important exception is Beaman and Dillon (2018) who look at how gender plays a role in information diffusion.

2. Experiment

2.1. **Setting and Sample.** Our study took place in Indonesia in 2015 and 2016. Despite being a developing country, Indonesia is quite active on social media, ranking third worldwide in the number of Facebook accounts, with 130 million[7] in 2020 (about half the population), and ranking eighth in the number of Twitter accounts, with over 10.6 million (about 6.4 percent of the population).[8] These Twitter users are active as well: in 2012, a study that linked individual tweets to their cities of origin found that Indonesia's capital, Jakarta, was the top city producing tweets anywhere in the world, narrowly exceeding Tokyo.[9]

The focus of the experiment was on improving immunization. Immunization was chosen as it was a government priority, as Indonesia was trying to improve immunization rates as part of its drive towards the Millennium Development Goals. A set of 550 tweets was developed in close coordination with the Ministry of Health that sought to improve information about immunization. The tweets included information about access to immunization (e.g., immunizations are free, available at government clinics, and so on); information about the importance of immunization (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 also identified a source (either a specific link or an organization's Twitter handle). All tweets were approved by the Ministry of Health, and all included a common hashtag, #AyoImunisasi ("Let's Immunize"). Each tweet was written in Indonesian, and two versions were prepared—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, to participate in our experiment. These "celebrities" come from a wide range of backgrounds, including pop 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 followers. While these celebrities primarily tweet about things pertaining to their main reason for fame, they also comment from time to time on public issues beyond their normal set of issues, often passing on sources or links, so tweets about immunization would

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

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

[9] https://semiocast.com/en/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US

not necessarily have been out of the ordinary.[10] We also recruited 9 organizations involved in public advocacy and/or health issues in Indonesia with a mean of 132,300 followers each with a total of 1.3 million followers.

In addition to the celebrities, we recruited 1032 ordinary citizens, whom we call "ordinary Joes and Janes". The role of the Joes/Janes will be to allow us to have essentially unimportant, everyday individuals compose tweets that are then retweeted by celebrities, which will be important for identification. These Joes/Janes consist primarily of university students at a variety of Indonesian universities. They are far more typical in their Twitter profiles, with a mean of 511 followers.

Every participant (both 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) randomize the content of the tweets from a large list of 549 immunization tweets approved by the Ministry of Health, and (3) has no scope for editing. Participants were given two choices: (1) the maximum number of tweets to authorize (13, 23, or 33), and (2) a writing style for the tweet (to better approximate their normal writing style), either formal or slang language.

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, 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, and hence ultimately an empirical question.[11]

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).[12] To ensure everything else about the tweets was balanced across these interventions, we also randomized the precise timing of tweets (which day and what time of day, matching the empirical frequency of local time-of-day of Indonesian tweets),

[10]For example, in the few months prior to our campaign, three of the celebrities in our sample (a musician, a TV personality, and well-known musician's agent) tweeted about the importance of breakfast, including a link to an article about the health benefits of children's breakfast; a well-known athlete tweeted about supporting education for young children; a musician tweeted in support of Asia Against Aids; and a well-known doctor tweeted about his support for the Indonesian anti-corruption agency.

[11]Online 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.

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

and the content of the tweet, (i.e. which tweet from our pre-prepared bank of approved tweets was tweeted by whom and when).[13] Celebrities were randomized into two phases, with half tweeting in the first phase and half in the second phase. In addition, towards the end of the last phase of our experiment, all tweets / retweets by a celebrity were then retweeted by a randomly selected number of Joes/Janes. These various randomizations allow us to identify the role of celebrity reach vs. endorsement, as well as the role of celebrity authorship and sourcing, as described in more detail in Section 3 below.

2.3. **Data.** We collected detailed data on relevant behavior on the Twitter platform via the Twitter Firehose and Twitter API. Before the experiment began, in early 2015, we collected an image of the publicly available Twitter network, including the list of followers of any celebrity participating in our study. This allows us to construct the follower network in Indonesia at baseline.

There are two main types of behaviors that people who see tweets can do, "likes" and "retweets". A *like* is a expression of approval of the tweet. A *retweet* is when someone who has seen a tweet rebroadcasts it to their entire follower network; this allows information to propagate over the Twitter network. There are two main differences between likes and retweets. First, retweets do not necessarily imply endorsement of the views of a tweet, whereas likes do. Second, while likes are visible (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. An example of a campaign tweet, with a source, is depicted in Figure 1.

For each the of the 672 total tweets that were 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, we also scraped all of this follower's followers and their liking and retweeting behavior.[14] For each of these events, we used the complete follower network (and followers' followers when the follower retweeted) from the baseline to construct the shortest path through which the tweet could have reached the user. 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. Online Appendix Table B.1 reports descriptive statistics.

[13]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.

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

3. Results

We begin in Section 3.1 by asking whether a celebrity's influence in diffusion on social media is due simply to their reach (the size of their network) or whether their involvement per se has an additional effect (the endorsement effect). We then seek to decompose the endorsement effect by studying choices celebrities can make when they speak, by separately identifying the role of celebrity authorship (Section 3.2) and citing of external sources (Section 3.3).

Before beginning, it is important to clarify that what we term the "endorsement effect" is asking whether the identity of the node in a network affects the probability of subsequent diffusion. If the endorsement effect is present, the value of a node is not simply its position in the network (i.e., its reach), but the specific label attached to it as well. Conceptually, this means that the endorsement effect nests a number of things. For example, that the celebrity is identified as passing on a message may affect subsequent retweeting and/or liking behavior either positively or negatively because: (a) this affects the message's perceived value; (b) it is socially desirable to demonstrate that one engages with celebrity content; (c) it is not socially desirable to retweet when one feels that others will do so in any case; (d) it may be more fun to retweet anything by a celebrity; (e) it is viewed as frivolous to retweet anything by a celebrity. And indeed this list is not necessarily exhaustive. But the key idea from each of these examples is that it is not simply the position of the node in the network that affects subsequent behavior, but the identity of who the node is *per se* that may affect diffusion: this is what we call the endorsement effect. We then decompose this into *involvement* (which does not discriminate between whether the celebrity passes on a message or writes it) and *authorship*.

3.1. **Identifying the Role of Celebrity Involvement.**

For involvement, our identification strategy exploits a particular feature of retweets in Twitter. A respondent j who sees a retweet observes two names: the name of the original writer of the tweet, and of the person whom j follows who retweeted it. Any names in between—say, a follower of the original writer who retweeted it to the person who retweeted it to j—are unobserved.

Consider a chain from a celebrity to some follower F_1 and then to some follower of this follower (who does not directly follow the celebrity) F_2. If the celebrity retweets the message by a Joe/Jane, and then this is retweeted by F_1, observe that 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 rather than retweeted it, this would be visible to the F_2. This is depicted in Figure 2.

By randomizing whether the message is originally tweeted by the celebrity, or instead originally tweeted by a Joe/Jane and then retweeted by the celebrity, we can identify the celebrity endorsement effect by looking at F_2's behavior.

We estimate, by Poisson regression, the equation

$$\mathrm{E}[y_{trcmp}|\mathbf{x}_{trcmp}] = \exp\left(\alpha \cdot \mathrm{Celeb}_{tcm} + \beta \cdot \log(\mathrm{Followers})_r + \omega_c + \omega_m + \omega_p\right) \tag{3.1}$$

where t indexes a tweet, r indexes a retweeter (i.e., an F_1 who retweeted the tweet t), c indexes a celebrity, 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 if visible to the F_2), and 0 if the celebrity retweeted a Joe/Jane (and hence her identity is not visible to the 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 to allow for arbitrary variance terms clustered at the original tweet (t) level. We control for the log number of followers of the F_1, and for dummies (ω_m) for the 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 chosen, etc., as well as phase fixed effects (ω_p).

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-level person making the decision to retweet.

Table 1 presents our results. As discussed above, 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 retweeting 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) by followers-of-followers (F_2s). The results are qualitatively similar when we look at likes or retweets 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 the message substantially increases the likelihood that the F_2 responds online.

The results also allow us to begin to decompose the reason for retweeting. Specifically, the estimates imply that 63-72%[15] of the retweets comes 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 D.1 of Appendix D.[16] We show that similar to celebrity effect, an organization being randomly assigned to compose a message rather than a Joe/Jane has a substantial endorsement effect of similar magnitude.

There are two main potential threats to identification here. First, when we look at F_2 agents, i.e., those who are at distance two from the celebrity of interest, 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, rather than retweeting it from a Joe/Jane. In equation (3.1), we always 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 there would be a bias in equation (3.1). But there may nevertheless be a *compositional* difference in which F_1s retweet it, which could potentially lead to selection bias in terms of which F_2s are more likely to see the retweet in the first place.

To address this issue, in the last phase of the experiment, we added an additional randomization. Specifically, we use the subset of Joe/Jane 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 $F_1 s$. For this sample, we can look at how *their* followers—that is, the followers of F_1 Joe/Jane's we exogenously forced to retweet a particular tweet—responded as we randomly vary whether the celebrity, an organization, or a Joe/Jane composes the message. We analyze this experiment by estimating equation (3.1) just as we did for the full sample of F_2s, but for this sample we have the advantage that whether an F_2 sees the tweet is guaranteed to be exogenous by construction.

Columns 2, 4, and 6 present the results. The point estimates are if anything somewhat larger than those in the full sample, and we cannot reject equality. Statistical significance is reduced somewhat in this restricted sample (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 whole sample is not leading to substantial bias.

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

[16] Recall that we only have 7 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 second potential confound comes from the fact that a retweet carries with it information about how many times the original tweet has been retweeted or liked as of the time the user views it (see Figure 1, which shows the number of retweets next to the arrow graphic and the number of likes next to the heart graphic). One may worry then that 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 us address this issue, because we randomly varied the number of Joes/Janes we forced to retweet a particular tweet. Appendix C, Table C.1 presents Poisson regressions of retweets and likes on the number of Joes/Janes that were forced to retweet a given celebrity's tweet or retweet (this is of course net of the forced Joe/Janes' behavior). 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.

3.2. **Decomposing Endorsement: Authorship vs. Pure Involvement.** The preceding analysis compared individuals who were effectively randomly blinded to whether a celebrity was or was not involved in the message composition and passing in order to estimate an endorsement effect. While involvement was through authorship in that case, we do not know if *involvement* per se or *authorship* per se mattered.

We can, however, use the same randomization – does the celebrity write themselves or retweet the message – to answer this question, but this time, looking at the direct followers of the celebrities themselves (i.e., the F_1s). For F_1s, they see the message either way, but the randomization changes whether it was directly authored by the celebrity or retweeted.

Table 2 presents the results of Poisson regressions at the tweet level. We restrict to direct followers of the celebrity (F_1 individuals), and estimate

$$\mathrm{E}[y_{tcmp}|\mathbf{x}_{tcmp}] = \exp\left(\alpha \cdot \mathrm{Celeb}_{tcm} + \omega_c + \omega_m + \omega_p\right). \tag{3.2}$$

We now have one observation per tweet, and look at the number of retweets/likes, retweets, or likes by F_1 agents who are distance 1 from the celebrity passing along the tweet. We continue to include celebrity (ω_c), phase (ω_p), and message-type (ω_m) fixed effects.

We find evidence that authorship is important. The estimates suggest that tweets authored by celebrities are 200 percent more likely to be retweeted/liked rate than those where the celebrity retweets a message (column 1, $p < 0.001$). In fact, an agent who observes a tweet composed from the celebrity rather than a retweet of a Joe/Jane is 120 percent more likely to like the tweet (column 2, $p < 0.001$) and 280 percent more likely to retweet the tweet (column 3, $p < 0.001$).

This fact allows us to further decompose the impacts of celebrity. The estimates here imply that 79% of the endorsement effect we estimated earlier come from the authorship per se. Combining the estimates here with those in the previous suggestion suggests that, on net, 56 percent of the celebrity effect comes from authorship, 14 percent from the endorsement

value, with the remainder attributable to the intrinsic interest of the message itself. These results suggest that celebrities can play an important role in the diffusion of public health messages, but to do so, they need to speak in their own voice.

3.3. **Citing Credible Sources.** Finally, we examine whether citing sources increases diffusive behavior. Every tweet in our databank was paired with a source, and 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.[17]

To examine this question, we re-estimate equation (3.2) at the F_1 level, but also add a variable that captures whether the tweet was randomly selected to include a source.[18] Table 3 presents the results. Columns 1-3 look at pooled likes and retweets, 4-6 look at retweets, and 7-9 look at likes. We also condition the regressions to the sample where the celebrity retweets the Joes/Janes (columns 2, 5, 8) and the celebrity directly tweets (columns 3, 6, 9).[19]

On average, pooling across all messaging configurations we find that source citation reduces the retweet and liking rate by 26.3 percent (-0.306 log points; $p = 0.051$) . This is particularly driven by reduced retweeting behavior (a decline of 27.2 percent or -0.318 log points, $p = 0.048$). Disaggregating across whether the celebrity composed and tweeted the message or retweeted a Joe/Jane or an organization, we find that the large reductions in retweet/like rates persist both when a Joe/Jane composed the message (a 50 percent or -0.553 log point decline, $p = 0.02$) or when the celebrity directly tweets the message (a 29.3 percent or -.235 log point decline, $p = 0.002$).

In sum, for both types of messages including a source in tweets ultimately depresses retweet rates and the extent of diffusion. This result—that sources depress retweet rates—may seem surprising, since one might expect that a sourced message may be more reliable. But recall the discussion in Section 2.2 suggesting that each feature of the message (e.g., originator identity, sourcing) could have ex-ante ambiguous effects on retweet rates. There

[17]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).

[18]Note that the number of observations is smaller here, because some tweets on topics deemed 'sensitive' by the Government always included a source, as we noted above. We restrict the analysis here to those tweets for which we randomized whether the source was included or not.

[19]These columns break out on these specific sub-samples we wish to focus on. Columns 1, 4, and 7 also include the cases where celebrities retweet organiztions.

are a number of possible explanations for this finding. One idea, which we emphasize 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 part in prior work (Chandrasekhar et al., 2018). We discuss in Online Appendix A how adding a source could potentially reduce the signaling value and possible lead to lower retweet rates (alongside a more general discussion of how celebrity origination, source citation, exposure, and specific content could all either increase or decrease retweet rates ex ante). Various other stories are possible as well. For instance, it is also possible that F_1s interpreted a sourced message as less authentic-sounding than an unsourced message, perhaps the information sounds less novel when sourced, or even if they click through to the source this dissuades them from returning and passing on the message itself. Regardless, the result suggests that adding explicit sources to celebrity messages does not necessarily increase diffusion.

4. Conclusion

We conducted a nationwide public health campaign in India, which consisted of a randomized controlled trial on Twitter involving 46 celebrities and organizations to promote immunization.

We examine what makes a campaign effective, decomposing celebrity influence through reach and endorsement effects and studying credible sourcing choices. Celebrity endorsement increases retweet rate by about 70 percent. Decomposing this between involvement per se and authorship, we find the vast majority (79%) of the endorsement effect stems from authorship rather than involvement. Source citation has an adverse effect.

Our findings shed some light on what effective design might look like: recruiting influential agents, like celebrities, to send self-authored messages in their own voice, without explicitly citing credible external sources.

References

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.

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.

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

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," .

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

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.

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*.

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

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

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.

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.

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.

King, G., B. Schneer, and A. White (2017): "How the news media activate public expression and influence national agendas," *Science*, 358, 776–780.

Kremer, M. and E. Miguel (2007): "The Illusion of Sustainability," *Quarterly Journal of Economics*, 122, 1007–1065.

Leskovec, J., L. A. Adamic, and B. A. Huberman (2007): "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, 1, 5.

Ryan, B. and N. C. Gross (1943): "The diffusion of hybrid seed corn in two Iowa communities." *Rural Sociology*, 8, 15.

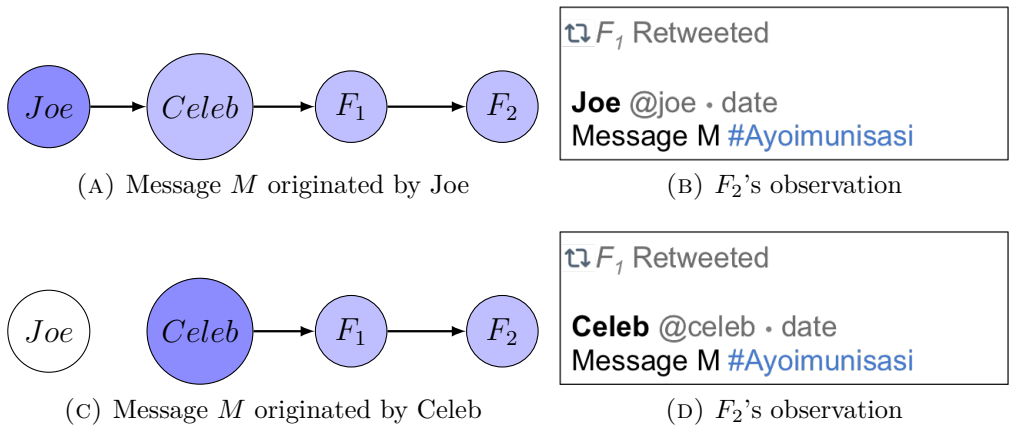
Figures

(A) Celebrity Tweet: casual with credibility boost

(B) Celebrity retweeting an Organization: casual with credibility boost

(C) Celebrity retweeting a Joe: formal without credibility boost

Figure 1. Sample tweets and retweets from the campaign



(A) Message M originated by Joe (B) F_2's observation

(C) Message M originated by Celeb (D) F_2's observation

FIGURE 2. Identification of the value of endorsement of celebrity involvement.

Tables

Table 1. Identifying the Role of Celebrity Endorsement

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Pooled	(3) Poisson # Retweets	(4) Poisson # Retweets	(5) Poisson # Likes	(6) Poisson # 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. Decomposing Endorsement: Authorship vs. Pure Involvement

Panel B: Measured through Composition measured by F_1 likes/retweets

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # 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

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, and content fixed effects.

TABLE 3. The Effects of Source Citation, measured by F_1 likes/retweets

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Pooled	(3) Poisson # Pooled	(4) Poisson # Retweets	(5) Poisson # Retweets	(6) Poisson # Retweets	(7) Poisson # Likes	(8) Poisson # Likes	(9) Poisson # Likes
Source cited	-0.306	-0.553	-0.235	-0.318	-0.694	-0.347	-0.277	-0.261	0.0104
	(0.157)	(0.248)	(0.109)	(0.161)	(0.297)	(0.113)	(0.183)	(0.236)	(0.248)
	[0.0513]	[0.0260]	[0.0319]	[0.0478]	[0.0195]	[0.00222]	[0.130]	[0.269]	[0.967]
Observations	492	170	131	492	170	131	492	170	131
Depvar Mean	3.644	2.635	7.305	3.644	2.635	7.305	3.644	2.635	7.305
Celeb RT Joe/Jane	✓	✓		✓	✓		✓	✓	
Celeb Direct	✓		✓	✓		✓	✓		✓

Notes: Standard errors (clustered at the celebrity/organization level) are reported in parentheses. p-values are reported in brackets. All regressions control for phase, celebrity fixed effects, content fixed effects, and condition on non-sensitive tweets.

ONLINE APPENDIX: NOT FOR PUBLICATION

Appendix A. Model

A.1. **Overview.** We study the decision by individuals on Twitter to pass on information to their followers by "retweeting" it. Before proceding to our empirical analysis, we begin by discussing a simple framework to think through how individuals make the decision to pass on information. The framework is standard, developed in Chandrasekhar, Golub, and Yang (2018) and also previously applied in Banerjee, Breza, Chandrasekhar, and Golub (2018).

In our framework, individuals pass on information for two reasons. First, individuals may care that others are informed about a topic. Second, as retweeting is intrinsicially a social activity, individuals can be motivated by how they are viewed by their followers. In this case, individuals may choose to retweet certain topics as a function of how the act of sharing the information changes how they are perceived by others. For example, individuals on Twitter may be trying to gather more followers, and it is plausible that people are more likely to keep following someone whom they believe is sharing high-quality information.

This second observation—that people may share information with a view to it affects how others perceive them—turns out to have subtle ramifications for how we think about a dissemination strategy. Whether information is more likely to spread more widely if originated by a celebrity or an ordinary Joe, or whether messages cite credible sources or simply consist of assertions, turn out to be ambiguous questions once we include the fact that these features of messages change the degree to which sharing the message provides information in equilibrium about the likely quality of the person deciding whether to share it.

In particular, the standard intuition is that more and credible information is simply better, and hence more likely to be retweeted. This comes from a standard model in which individuals only base their decisions to pass on information based on the first factor, namely the quality of that information. In this case, if a message has more credibility and has a verified source, then more retweeting should happen. This generates an intuition that, for instance, sourced tweets or celebrity tweets should be retweeted more.

However, when we consider the fact that retweeting has a social component—that individuals certainly care about how they are perceived and that is likely a key component of their motivation to retweet—we see that these conclusions change. Assume that an individual F follows an originator of a tweet, o. Suppose that F is more willing to pass on information if he is more certain about the state of the world. Also assume that F can be one of two private types: a high type (greater ability or social consciousness for the sake of discussion) and a low type. Individuals desire to be perceived of as a high type by their followers, so part of the motivation to retweet is for this social perception payoff. It is commonly known that high types are better able to assess the state of the world rather than low types (i.e., imagine that in addition to the tweet, individual F gets a private signal as to the state of

the world, and the high types' signal is more informative). When F sees a tweet by o, he needs to glean the state of the world using both the tweet and his own private signal, and decide whether or not to retweet.

To illustrate ideas, let us compare the case where o's tweet contains no source versus cites a credible source about the topic. Inclusion of a source has multiple effects. First, the source citation should make the state of the world even more evident. This should encourage retweeting through increasing certainty. Second, and more subtly, if social perception is important enough, source citation can have a discouraging effect on retweeting. Specifically, if a source makes it very clear what is true, then there is no room for signaling remaining: high types are no better able to assess things than low types and therefore ability does not really matter. We show below that which effect dominates on net—the increased direct effect of the source on quality, or the fact that the source decreases the ability of F to use the tweet to signal quality—turns out to be ambiguous.

To show this more formally, we adapt the endogenous communication model developed by Chandrasekhar, Golub, and Yang (2018) to our context of retweeting on Twitter (see also Banerjee et al. (2018) for another such prior application of this model). Such image concerns have also been looked at both theoretically and empirically in both Bursztyn and Jensen (2015); Bursztyn, Egorov, and Jensen (2017) who study whether peer perceptions inhibit the seeking of education. We look at individuals who have payoffs from passing on information and who are concerned with social perception as well the direct value of the information they pass. We show how that sourcing, originator identity, exposure, and content all can have ambiguous effects on the amount of retweeting, and explore when we might expect which policies to work well.

It is important to note that we are not claiming of course that these are the only motives for retweeting. After all there can be more mundane motivations: it is just more fun to retweet anything by a celebrity, it is just frivolous to retweet anything by a celebrity, one likes to retweet something that he/she anticipates will not be otherwise widely spread, among other explanations. But without hardcoding anything else into the model, in the simplest interpretation of dynamics on Twitter, we can demonstrate and motivate why the questions we study are ultimately empirical issues.

A.2. **Setup.**

A.2.1. *Environment.* The state of the world is given by $\eta \in \{0, 1\}$, with each state equally likely. There is an originator o (she) who writes an initial message about the idea with probability $q \in (0, 1]$, which is received by her follower F (he). With probability $1 - q$ nothing happens. The message is a binary signal about the state of the world, which is

accurate with probability α, i.e.

$$m = \begin{cases} \eta & \text{w.p. } \alpha \geq \frac{1}{2} \\ 1-\eta & \text{o.w.} \end{cases}.$$

The message may or may not cite a source, designated by $z \in \{S, NS\}$ respectively. We allow the quality of the signal to depend on source, so $\alpha = \alpha_z$, discussed below.

Further, there are two types of originators: ordinary Janes/Joes and celebrities, given by $o \in \{J, C\}$ respectively. We allow the quality of the signal to depend on originator, so $\alpha = \alpha_o$, discussed below.

Finally, followers come in two varieties: $\theta \in \{H, L\}$ represents F's privately known type, and one's type is drawn with equal odds. High types have better private information about the state of the world. This can represent ability in a loose way such as intelligence, social accumen, taste-making ability, or any trait which allows F to better discern the state of the world if he is of type H rather than L. We model this by supposing that F draws an auxiliary signal, x, with $x = \eta$ with probability π_θ and $x = 1-\eta$ with probability $1-\pi_\theta$. We assume $\pi_H \geq \pi_L$ which reflects that H-types can better discern whether the idea is valuable. As discussed below, it is socially desirable to be perceived as $\theta = H$.

This environment captures our basic experimental setting. We randomly vary originator $o \in \{J, C\}$ and whether the message is sourced, $z \in \{S, NS\}$.

A.2.2. *Bayesian Updating.* F is assumed to be Bayesian. Let $\alpha = \alpha_{o,z}$ be the quality of the signal depending on originator and source. Therefore given message m and private signal x, we can compute the likelihood ratio that F believes the state of the world being good versus bad as

$$\begin{aligned} LR\left(\eta | m, x; o, z, \theta\right) &= \frac{\mathrm{P}\left(\eta = 1 | m, x\right)}{\mathrm{P}\left(\eta = 0 | m, x\right)} = \frac{\mathrm{P}\left(m, x | \eta = 1\right)}{\mathrm{P}\left(m, x | \eta = 0\right)} \\ &= \left(\frac{\alpha_{o,z}}{1-\alpha_{o,z}}\right)^m \left(\frac{1-\alpha_{o,z}}{\alpha_{o,z}}\right)^{1-m} \left(\frac{\pi_\theta}{1-\pi_\theta}\right)^x \left(\frac{1-\pi_\theta}{\pi_\theta}\right)^{1-x}. \end{aligned}$$

Note that as α or π tend to 1 or $\frac{1}{2}$, the likelihood ratio tends to $+\infty$ (the signal reveals the state) or 1 (the signal has no content), respectively.

A.2.3. *Payoffs.* The utility of F depends on two components. The first is the instrumental payoff: it is a payoff from retweeting when the state of the world is more clear: that is when $LR\left(\eta\right)$ is more extreme. Thus we assume that the instrumental payoff when you do not retweet, i.e., when $r = 0$, is 0 and when you do retweet, i.e., $r = 1$, is $\varphi\left(LR\left(\eta | m, x; o, z, \theta\right)\right)$ for some smooth increasing in distance function from 1, $\varphi\left(\cdot\right)$. What this captures is that there is more instrumental value in passing on a message the greater certainty in the state

of the world. For instance if we set

$$\varphi(x) = f\left(\left|\frac{x}{1+x} - 1\right|\right)$$

for a smooth increasing function $f(\cdot)$ on $\left[0, \frac{1}{2}\right]$, the instrumental value is a monotone function in the probability the state of the world is high, but other functions φ will also work.[20] Further, due to taste or cost heterogeneity, there is a shock ϵ to the instrumental payoff of retweeting, where ϵ is a mean-zero random variable drawn from a continuous CDF with full support, such as the logit CDF $\Lambda(\cdot)$. Altogether, the instrumental payoff V^r is given by

$$V^r = \begin{cases} \varphi(LR(\eta|m,x;o,z,\theta)) - \epsilon & \text{if } r = 1 \\ 0 & \text{if } r = 0. \end{cases}$$

The second is the social perception payoff. Specifically F is concerned with the posterior that his followers have about his type given his decision to retweet: $\psi(\mathrm{P}(\theta = H|r))$ where $\psi(\cdot)$ is a monotonically increasing function. The perception in equilibrium is simply a function of the retweet decision itself. The idea here is that someone who is more able is more likely to be able to discern valuable topics and therefore the equilibrium decision to retweet itself has a signaling component. [21]

F's total utility is given by

$$U(r|m,x) = \underbrace{V^r}_{\text{instrumental}} + \underbrace{\lambda\psi(\mathrm{P}(\theta = H|r))}_{\text{perception}}$$

where $\lambda \geq 0$ is a parameter that tunes the strength of the perception payoff. [22]

Correspondingly, the marginal utility of choosing $r = 1$ versus $r = 0$ is given by

$$MU(r|m,x) = \underbrace{\varphi(LR(\eta|m,x;o,z,\theta)) - \epsilon}_{\text{change in instrumental}} + \underbrace{\lambda\Delta_r\psi(\mathrm{P}(\theta = H|r))}_{\text{change in perception}}.$$

Let $Q_H(\cdot)$ be the CDF of $\varphi(LR(\eta|m,x;o,z,H)) - \epsilon$ and $Q_L(\cdot)$ be the CDF of $\varphi(LR(\eta|m,x;o,z,L)) - \epsilon$.[23] It immediately follows that $Q_H \succ_{\text{FOSD}} Q_L$. This can be seen by inspection, where the likelihood ratio under type H first order stochastically dominates that of type L when $\eta = 1$ and the inverse of the ratio first order stochastically dominates when $\eta = 0$. It will be useful

[20] To see this, note that

$$\varphi(LR(\eta|m,x;o,z,\theta)) = f\left(\left|\frac{LR}{1+LR} - 1\right|\right) = f\left(\left|\mathrm{P}(\eta = 1|m,x;o,z,\theta) - \frac{1}{2}\right|\right)$$

which is just a smooth function of distance from pure uncertainty of a belief of $\frac{1}{2}$.

[21] For simplicity we abstract from F's followers interpretation of m and their own subsequent private signals. The reason is that we can demonstrate interesting non-monotonicities in retweeting behavior as a function of message quality without such additions, which would only serve to complicate matters.

[22] While λ could be absorbed into $\psi(\cdot)$, it is useful for exposition to keep it separate.

[23] This holds fixed o and z.

below to denote by G_θ the complementary CDF, $G_\theta := 1 - Q_\theta$, i.e., $G_\theta(v)$ is the fraction of types θ with a (net-of-costs) instrumental value of passing greater than or equal to v.

A.3. **Analysis.** F decides to retweet if and only if $MU(r|m,x) \geq 0$. This decision trades off two components. On the one hand is the relative instrumental benefit (or cost) of pasing on the message, which is an increasing function of the likelihood that the state of the world $\eta = 1$, and is given by $\varphi(LR(m,x|o,z,\theta))$. On the other hand, retweeting itself changes the perception of F by his followers, given by$\Delta_r \psi(\mathrm{P}(\theta = H|r))$, and so the (equilibrium) relative gain/loss of reputation must be taken into account.

The model is formally characterized in Proposition 1 of Chandrasekhar et al. (2018), and we refer the interested reader to that paper for proofs. Chandrasekhar et al. (2018) show that under the above assumptions, an equilibrium exists, and will be in cutoff strategies where F chooses to retweet if and only if $\varphi(LR(\eta|m,x;o,z,\theta)) - \epsilon \geq v$ for some v. An equilibrium is characterized by a cutoff $\underline{v} < 0$, which is used by all Fs irrespective of type θ, where it is the solution to

$$\underline{v} = \lambda\psi(\mathrm{P}(\theta = H|r=0)) - \lambda\psi(\mathrm{P}(\theta = H|r=1)).$$

Here the equilibrium posteriors are determined by:

$$\frac{\mathrm{P}(\theta = H|r=0)}{1-\mathrm{P}(\theta = H|r=0)} = \frac{1-qG_H(v)}{1-qG_L(v)} \text{ and } \frac{\mathrm{P}(\theta = H|r=1)}{1-\mathrm{P}(\theta = H|r=1)} = \frac{G_H(v)}{G_L(v)}.$$

The intuition for the equilibrium is as follows. First, note that F's type does not matter for the decision he makes conditional on the draw v. That is, while θ affects the distribution of the instrumental value, once F knows his instrumental value, he is trading off that against the change in reputation due to his behavior. Therefore the cutoff (in utility space) will not depend on θ's type.

At the cutoff $\underline{v}$ in equilibrium the marginal benefit of retweeting (which is a way to gain reputation by being viewed as more likely to be a high type) must be equal to the marginal cost of retweeting (which in this case is the instrumental benefit of passing the information relative to the stochastic cost). The reason $\underline{v} < 0$ is because here retweeting is a signal of being the high type, and therefore some low types will opt into retweeting despite having a negative net instrumental cost.

Holding fixed o, z as we have been doing above, we can compute the retweeting share in equilibrium:

$$\frac{1}{2}G_H(\underline{v}) + \frac{1}{2}G_L(\underline{v}).$$

We can also look at several contrasting situations. In the first, assume that $\lambda = 0$ with the same setup as above, so there is no interest in social concerns. Then only positive

instrumental values are retweeted, so the share retweeting is given by

$$\frac{1}{2}G_H(0) + \frac{1}{2}G_L(0).$$

Clearly the retweeting share is lower than when there is also a signaling motive, which featured an equilibrium cutoff $\underline{v} < 0$.

A second contrasting situation is one in which while individuals would potentially care about signaling, neither party is better at discerning the state of the world. That is, $\widehat{G}_H = \widehat{G}_L =: \widehat{G}$. In this case the share retweeted again is only determined by positive instrumental values and therefore is given by

$$\widehat{G}(0).$$

Whether $\widehat{G}(0) \lesseqgtr \frac{1}{2}G_H(\underline{v}) + \frac{1}{2}G_L(\underline{v})$ depends on how $\widehat{G}$ compares to G_H and G_L.

A subtle feature of the model is the fact that the retweet share is not necessarily monotonically increasing in the quality of the message, α. Intuitively, there are two effects of increasing α of retweeting. First, as α increases, the message becomes more informative. This increases the instrumental value of retweeting, and hence retweeting increases with α. Second, as α increases, the m signal becomes more informative relative to the private x signal. This makes the act of retweeting more about m than x, and hence lowers the signaling value of retweeting. Indeed, in the limit where $\alpha = 1$, there is no signaling value whatsoever. Thus, the signaling effect leads to a reduction in the amount of retweeting as α increases. Which effect dominates depends on parameters, and as we show now, in fact the effect of α on retweeting can be non-monotonic under some configurations of parameters.

Figure A.1 presents simulation results to further illustrate these intuitions. First consider the case when there is no reputation considerations ($\lambda = 0$). In this case, as the message's quality increases, the share retweeting must increase clearly because the value of information on average increases.

Next let us consider the case where neither H nor L are particularly able types, with $\pi_H = 0.53$ and $\pi_L = 0.5$. In this case, there is limited scope for signaling because the priors are quite poor: both types heavily lean on the message's signal m rather than their personal signals x. As such, like in the case with $\lambda = 0$, the quality of the message increases the share to retweet.

In contrast, consider the case where both types are expert, but H-types are somewhat better ($\pi_H = 0.95$, $\pi_L = 0.9$). In this case, with low α, since the predominant component of instrumental value comes from type to begin with, and because high types are much more likely to receive correct signals than low types but both have typically good signals about the state of the world (so m and x will agree), many more L types will also find it worthwhile to essentially "pool" with H types despite negative instrumental values due to reputation concerns. This leads to a monotonic decline in the retweet rate as α increases, since there

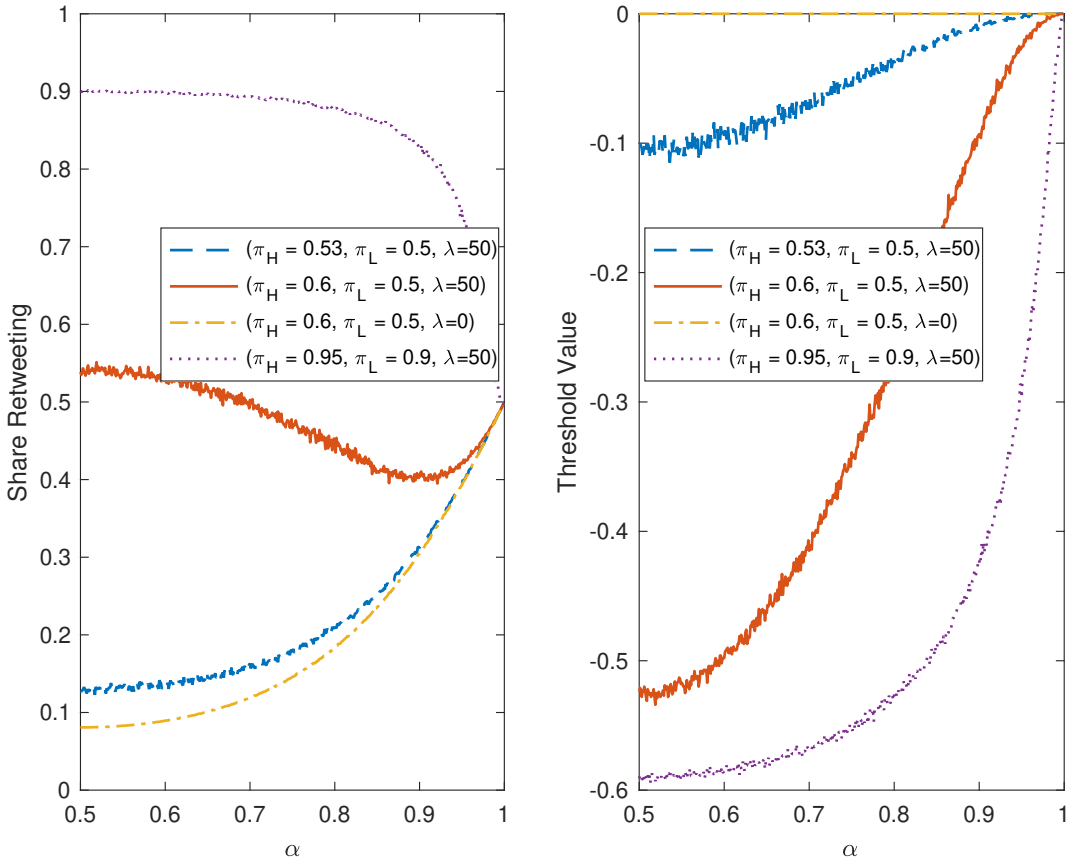


FIGURE A.1. Retweet share for various combinations of (π_H, π_L, λ).

is increasing reliance on the m-signal. What this means in practice is that it is possible to improve the quality of the message and yet reduce the overall share of retweeting, contrary to the naive intuition without a social perception payoff component.

The final case we show is the intermediate one, with $\pi_H = 0.65$ and $\pi_L = 0.5$. The signaling effect at this parameter level dominates initially, and hence increasing α initially decreases retweeting, but then eventually is dwarfed by the instrumental effect as the m-signal is considerably better than the gap in quality for the x-signal across types.

The fact that the relationship between α and retweeting is non-monotonic means that it is possible that mild increases in informativeness can reduce retweeting whereas dramatic increases in informativeness can increase it.

A.4. **Application to Experiment.** In what follows, we use the above framework consider possible implications of our experimental variations, i.e., (1) whether the originator is a celebrity or a Jane/Joe and (2) whether the tweet has a source or not.

A.4.1. *Celebrity versus Jane/Joe.* Celebrities and Joes can vary in the quality of their messaging. As such, we consider α_C versus α_J. Ex ante it may be possible for these to have any relationship, though we might think that celebrities tend to generate higher-quality signals. This could be, for instance, because celebrities' messages reach many more individuals and therefore they need to be more cautious in their messaging, or could be because they have better access to information in general.

Assuming $\alpha_C \geq \alpha_J$ and since $\eta = 1$ for an experimental topic (since all our messages are sent about true beneficial effects of immunization),

$$\mathrm{E}_{m,x}\left[\varphi\left(LR\left(\eta|m,x;C,\theta\right)\right)\right] \geq \mathrm{E}_{m,x}\left[\varphi\left(LR\left(\eta|m,x;J,\theta\right)\right)\right]$$

and therefore the distribution of instrumental payoffs $Q_{C,\theta} \succ Q_{J,\theta}$ for each θ. Note that this depends both on the originator and the type of the individual.

To see the effect, consider the case when $\alpha_C \to 1$. In this case, following the intuition discussed above, $Q_{C,H} \to Q_{C,L}$ and let $\widehat{Q}_C(\cdot)$ be the resulting CDF of the instrumental value, so there is nothing to signal at all. Thus $\underline{v}^C = 0$ and so anyone with any positive instrumental value immediately retweets. In contrast, with Joes, as above there is some negative $\underline{v}^J < 0$ that sets the equilibrium.

Consequently, the retweeting share is given by

- $\widehat{G}_C(0)$ under Celebrity origination and
- $\frac{1}{2}G_{J,H}\left(\underline{v}^J\right) + \frac{1}{2}G_{J,L}\left(\underline{v}^J\right)$ under Joe origination.

Notice that it is not clear which dominates. On the one hand, since $\eta = 1$ is essentially revealed as $\alpha_C \to 1$, $\widehat{G}_C$ has a higher mean than $G_{J,\theta}$ for either θ. On the other hand, the cutoff $\underline{v}^J$ can be considerably below 0 making the point of evaluating the $G_{J,\theta}$ CDFs at a lower point. This is because the likelihood ratio distribution of knowing that we are in a "good" world is not the same under celebrities (where it is substantially more likely) and Joes/Janes (where it is less likely, but there is a signaling effect reason to retweet).

Remark **1.** *The total endorsement effect we identify in the experiment can be thought of being comprised of (a) a shift in instrumental value and (b) a shift in the threshold to retweet due to the signaling effect. To see this*

$$\begin{aligned}\frac{1}{2}\left[\widehat{G}_C(0) - G_{J,H}\left(\underline{v}^J\right)\right] + \frac{1}{2}\left[\widehat{G}_C(0) - G_{J,L}\left(\underline{v}^J\right)\right] &= \frac{1}{2}\left[\widehat{G}_C(0) - G_{J,H}(0)\right] + \frac{1}{2}\left[G_{J,z,H}(0) - G_{J,H}\left(\underline{v}^J\right)\right] \\ &\quad + \frac{1}{2}\left[\widehat{G}_C(0) - G_{J,L}(0)\right] + \frac{1}{2}\left[G_{J,z,L}(0) - G_{J,L}\left(\underline{v}^J\right)\right].\end{aligned}$$

In this expression, the $\widehat{G}_C(0) - G_{J,\theta}(0)$ term measures how for a given cutoff of 0, the amount of retweets increases when a message is originated by the celebrity, and the the $G_{J,\theta}(0) - G_{J,\theta}\left(\underline{v}^J\right)$ term measures the change in the share of retweets when we move the cutoff to the left due to the signaling effect, holding the distribution fixed. When the signaling

impetus is dominant, this second term can overtake the prior term, making even a celebrity originator generate a lower volume of retweets.

A.4.2. *Sourcing.* In this case, holding originator fixed, we study the effect of adding a source. The analysis is identical to the case with celebrities. Ex-ante it seems reasonable to model sourcing as having direct positive effect on the likelihood of the signal being true: $\alpha_S \geq \alpha_{NS}$. Consequently

$$\mathrm{E}_{m,x}\left[\varphi\left(LR\left(\eta|m,x;S,\theta\right)\right)\right] \geq \mathrm{E}_{m,x}\left[\varphi\left(LR\left(\eta|m,x;NS,\theta\right)\right)\right].$$

This comes from the fact that a sourced tweet is just more likely to be right, so the likelihood ratio will be higher in distribution so for every originator and type of F, sourced tweets have more value in distribution so $Q_S \succ Q_{NS}$.

Again, if we assume sources are fully revealing $\alpha_S \to 1$ but without a source we have $\underline{v}^{NS} < 0$. Retweeting shares are given by $\widehat{G}_S\left(0\right)$ and $\frac{1}{2}G_{NS,H}\left(\underline{v}^{NS}\right) + \frac{1}{2}G_{NS,L}\left(\underline{v}^{NS}\right)$ under sourcing and no sourcing, respectively.

Crucially, even assuming sources are intrinsically good, retweeting can be reduced. This comes from the fact that the perception payoff effect can simply outweigh the gains in quality. If there is a source there is nothing to signal, whereas if there is no source F has a signaling motivation that is traded off against quality.

Remark **2.** *A natural question to ask is whether the since the arguments for celebrity versus Joe/Jane and sourced versus unsourced are identical, if anything seemingly relabeling, then the effects of sourced messaging and celebrity origination must have the same sign. But more careful reflection demonstrates that this is not true. Recall that retweeting share can be non-monotonic in α in this model. That is, given an initial α, a move to some $\alpha' > \alpha$ can lead to a decline in retweeting share and whether this is the case can depend on (π_H, π_L, λ). Concretely, recall the case of $(\pi_H = 0.65, \pi_L = 0.5, \lambda = 50)$ in Figure A.1 where the retweet share is non-monotonic with α. Thus, the increase due to a celebrity versus the increase due to adding a source need not be the same and in fact can generate different signs on retweeting behavior.*

Appendix B. Summary Statistics

Table B.1. User summary stats

	mean	obs
Followers of celebrities	262648	37
Followers of organizations	145300	9
Followers of Joes/Janes	574	134
Followers of forced Joes/Janes	502	898
Followers of celeb followers	1379	1073

Table B.2. Balance Check

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) OLS	(6) OLS
VARIABLES	Facts	Importance Info	Access Info	Myth-busting Facts	Other Facts	Source cited
Celeb writes and tweets	0.0302	-0.0580	0.0278	0.0139	0.0163	-0.0184
	(0.0449)	(0.0430)	(0.0318)	(0.0451)	(0.0255)	(0.0469)
	[0.501]	[0.178]	[0.383]	[0.758]	[0.524]	[0.695]
Observations	451	451	451	451	451	451
Phase control	✓	✓	✓	✓	✓	✓
Log #followers control	✓	✓	✓	✓	✓	✓
Message style control	✓	✓	✓	✓	✓	✓

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, formality, and exception status.

Appendix C. Does RT count affect retweeting?

Table C.1. Impact of No. of Forced Joe RTs on F_2 and F_1 likes/retweets

	(1) F2 Poisson	(2) F1 Poisson
VARIABLES	# Retweets	# Retweets
5 Forced Joe RTs assigned	0.0399	0.444
	(0.346)	(0.388)
	[0.908]	[0.252]
10 Forced Joe RTs assigned	0.244	0.0395
	(0.414)	(0.440)
	[0.556]	[0.928]
15 Forced Joe RTs assigned	0.256	0.207
	(0.407)	(0.359)
	[0.529]	[0.565]
Observations	505	184
Phase Control	✓	✓
Log #followers control	✓	✓
Message style control	✓	✓
Depvar Mean	0.184	2.707
1 Forced Joe RT assigned log mean	-2.331	0.870

Notes: Robust standard errors are reported in parentheses. p-values are reported in brackets.

Appendix D. Effect of celebrity retweeting organizations

Table D.1. Identifying the Role of Celebrity and Organization Endorsement

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # Likes
Celeb writes and tweets	0.423	0.421	0.574
	(0.182)	(0.179)	(0.520)
	[0.0201]	[0.0185]	[0.269]
Org writes and Celeb retweets	0.564	0.600	0.255
	(0.221)	(0.258)	(0.520)
	[0.0107]	[0.0200]	[0.624]
Observations	1,791	1,791	1,791
Joe writes mean	0.0417	0.0343	0.00745

Notes: Standard errors (clustered at the original tweet level) are reported in parentheses. p-values are reported in brackets. The sample conditions on tweets that are not sensitive and includes tweets originated by Joes, organizations, and celebrities. All regressions control for phase, celebrity fixed effects, and content fixed effects.