Appendix A. Main Analysis

Table A.1. Estimating the average value of celebrity involvement using followers-of-followers' (F_2) behavior

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
Phase control	✓	✓	✓	✓	✓	✓
Log #followers control	✓	✓	✓	✓	✓	✓
Message style control	✓	✓	✓	✓	✓	✓
Joe writes mean	0.0417	0.00915	0.0343	0.00686	0.00745	0.00229
Forced Joes 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 A.2. Value of Celeb Endorsement through Composition measured by F1 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
Phase control	✓	✓	✓
Log #followers control	✓	✓	✓
Message style control	✓	✓	✓
Joe writes and Celeb retweets mean	2.058	1.045	1.013

Notes: Sample conditions on all tweets originated by Joes/Janes or celebrities. All regressions control for phase, celebrity fixed effects, content fixed effects. Standard errors (clustered at the celebrity/organization level) are reported in parentheses.

TABLE A.3. Value of Endorsement through Source Citation measured by F1 likes/retweets

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Pooled	(3) Poisson # Pooled	(4) Poisson # Pooled	(5) Poisson # Retweets	(6) Poisson # Retweets	(7) Poisson # Retweets	(8) Poisson # Retweets	(9) Poisson # Likes	(10) Poisson # Likes	(11) Poisson # Likes	(12) Poisson # Likes
Source cited	-0.306	-0.553	-0.235	-0.0365	-0.318	-0.694	-0.347	0.0946	-0.277	-0.261	0.0104	-0.239
	(0.157)	(0.248)	(0.109)	(0.207)	(0.161)	(0.297)	(0.113)	(0.186)	(0.183)	(0.236)	(0.248)	(0.297)
	[0.0513]	[0.0260]	[0.0319]	[0.860]	[0.0478]	[0.0195]	[0.00222]	[0.612]	[0.130]	[0.269]	[0.967]	[0.421]
Observations	492	170	131	191	492	170	131	191	492	170	131	191
Depvar Mean	3.644	2.635	7.305	2.031	3.644	2.635	7.305	2.031	3.644	2.635	7.305	2.031
Celeb RT Joe/Jane		✓				✓				✓		
Celeb RT Org				✓				✓				✓
Celeb Direct			✓				✓				✓	

Notes: Sample conditions on 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 A.4. Did people offline hear about the campaign?

VARIABLES	(1) Logit Heard of *#Ayoimunisasi*	(2) Heard from Twitter	(3) Poisson # of times heard from twitter
Std. Exposure to tweets	0.219	0.108	0.125
	(0.114)	(0.0674)	(0.0531)
	[0.0560]	[0.110]	[0.0183]
	{.075}	{.19}	{.12}
Observations	2,154	2,404	2,441
Potential exposure control	✓	✓	✓
Double Post-LASSO	✓	✓	✓
Depvar Mean	0.0775	0.181	0.322

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 A.5. Did people offline increase knowledge?

VARIABLES	(1) Logit Domestic	(2) Correct on Substitutes	(3) Correct on Side-effects	(4) Logit Free
Std. Exposure to tweets	0.101	-0.0479	0.0253	0.0548
	(0.0654)	(0.0668)	(0.0638)	(0.0729)
	[0.122]	[0.473]	[0.692]	[0.452]
	{.082}	{.77}	{.62}	{.656}
Observations	2,421	2,430	2,439	2,416
Potential exposure control	✓	✓	✓	✓
Double Post-LASSO	✓	✓	✓	✓
Depvar Mean	0.575	0.527	0.487	0.677

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 A.6. Communication With and Behavior of Neighbors, Friends, and Relatives

Panel A: Communication: Knowledge of immunization behavior of others

VARIABLES	(1) Neighbor	(2) Friend	(3) 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

VARIABLES	(1) Neighbor	(2) Friend	(3) Relative	(4) Own
Std. Exposure to tweets	0.194	0.246	0.140	-0.00217
	(0.107)	(0.0955)	(0.0997)	(0.136)
	[0.0707]	[0.00994]	[0.159]	[0.987]
	{.133}	{.071}	{.06}	{.8260000000000001}
Observations	682	682	682	621
Potential exposure control	✓	✓	✓	✓
Double post-LASSO	✓	✓	✓	✓
Depvar Mean	0.356	0.353	0.314	0.485

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.

Appendix B. Model

B.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 intrinsically 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 how 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/Jane, 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 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.

B.2. **Setup.**

B.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, $z \in \{S, NS\}$.

B.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.

B.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.[23] 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. [24]

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. [25]

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 $\varphi(LR(\eta|m,x;o,z,L)) - \epsilon$.[26] 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

[23] 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 $\frac{1}{2}$.

[24] 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.

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

[26] 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.

B.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 passing 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 B.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

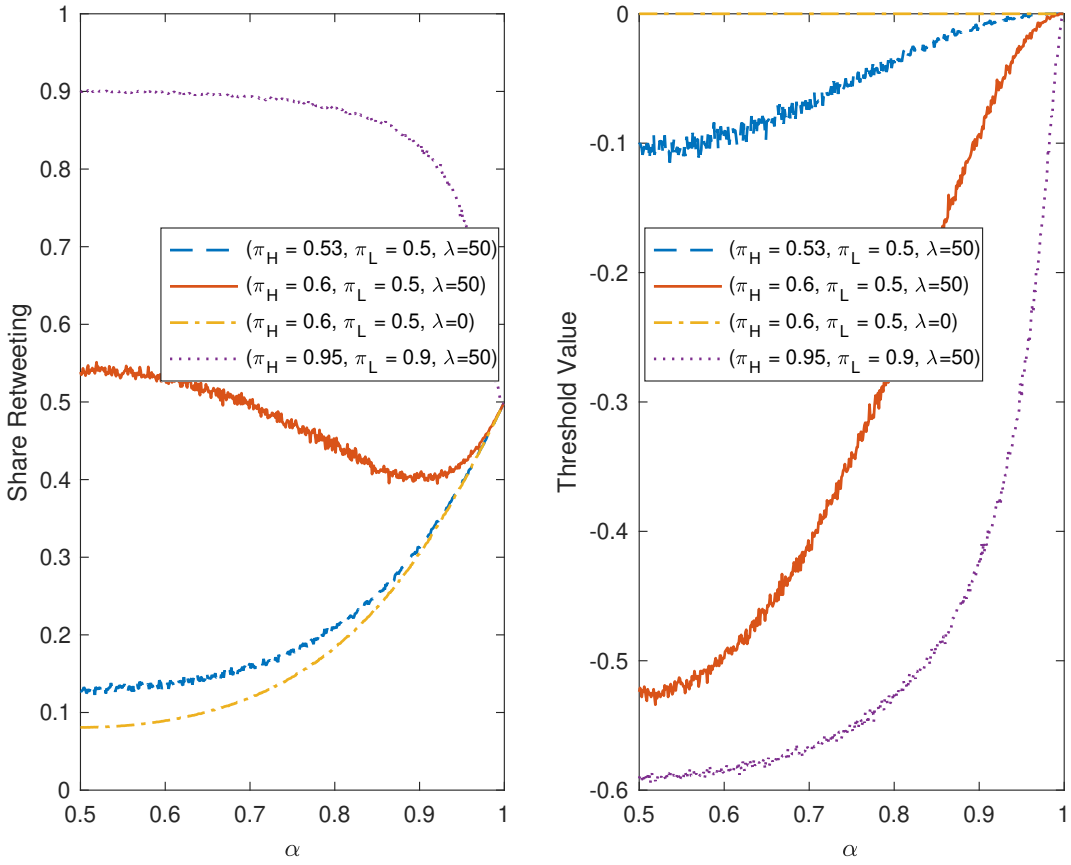


FIGURE B.1. Retweet share for various combinations of (π_H, π_L, λ).

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.

B.4. **Application to Experiment.** In what follows, we use the above framework to 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.

B.4.1. *Celebrity versus Jane/Joe.* Celebrities and Joes/Janes 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 it 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/Janes, 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 $G_{J,\theta}(0) - G_{J,\theta}\left(\underline{v}^J\right)$ 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.

B.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 $Q_S \succ Q_{NS}$.

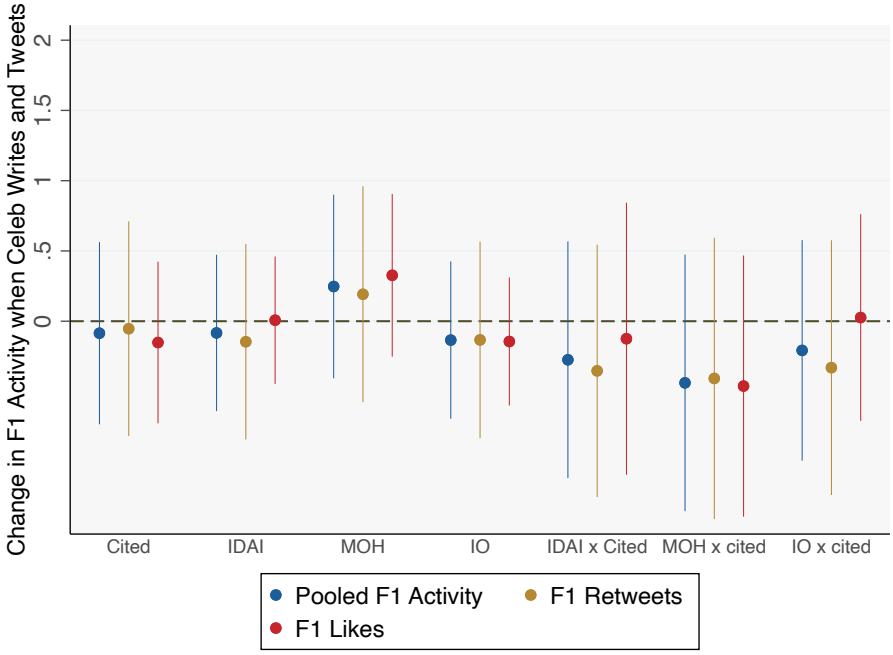
Again, if we assume sources are fully revealing $\alpha_S \to 1$ but without a source we have $\underline{v}^{NS} < 0$. Retweeting shares are given by $\widehat{G}_S\left(0\right)$ and $\frac{1}{2}G_{NS,H}\left(\underline{v}^{NS}\right) + \frac{1}{2}G_{NS,L}\left(\underline{v}^{NS}\right)$ under sourcing 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, 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 B.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.*

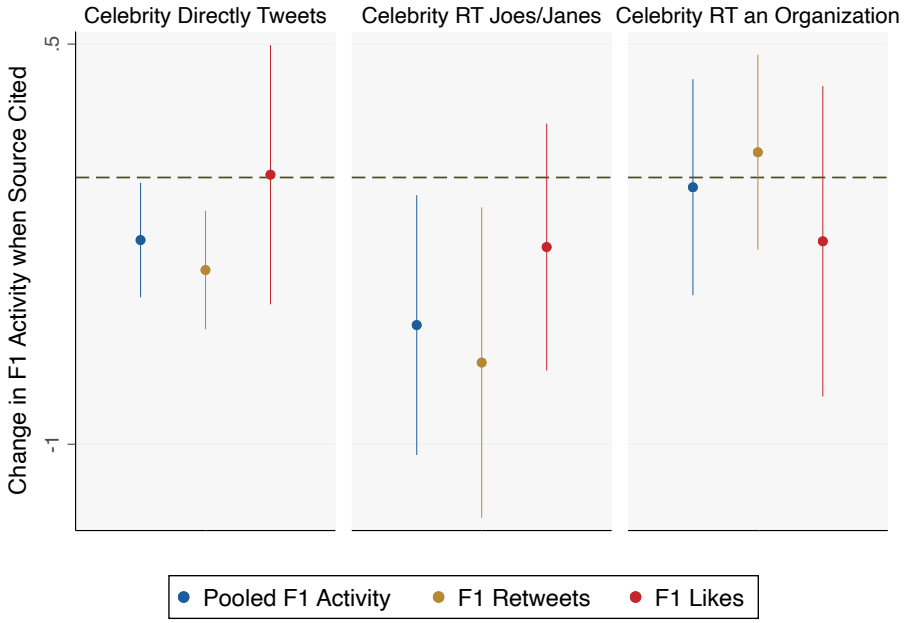
Appendix C. Source Heterogeneity

Figure C.1. Value of Endorsement through Source Citation measured by F1 likes/retweets with Heterogeneity by Source Cited



Note: Figure depicts regression estimates with their 95% confidence intervals. Equation model estimated is (2.2). All columns include fixed effects for phase and condition on non-exception tweets. The omitted categories are other types of sources, and its interaction with source being cited. These coefficients are also reported in regression table C.1.

FIGURE C.2. Value of Endorsement through Source Citation measured by F1 likes/retweets with Heterogeneity by how the Source is Cited



Note: Figure depicts regression estimates with their 95% confidence intervals. Equation model estimated is (2.2). Sample conditions on non-sensitive tweets. All regressions control for phase, celebrity fixed effects, content fixed effects. Standard errors are clustered at the celebrity/organization level. These coefficients are also reported in regression table A.3.

TABLE C.1. Value of Endorsement through Source Citation measured by F1 likes/retweets with Heterogeneity by Source Cited

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # Likes
Source cited	-0.0854	-0.0533	-0.152
	(0.331)	(0.390)	(0.293)
	[0.796]	[0.891]	[0.605]
IDAI	-0.0834	-0.146	0.00738
	(0.284)	(0.355)	(0.231)
	[0.769]	[0.680]	[0.975]
MOH	0.247	0.192	0.327
	(0.333)	(0.392)	(0.295)
	[0.459]	[0.625]	[0.269]
Intern Org	-0.134	-0.133	-0.144
	(0.285)	(0.357)	(0.232)
	[0.638]	[0.710]	[0.534]
IDAI x Cited	-0.274	-0.353	-0.125
	(0.430)	(0.458)	(0.494)
	[0.523]	[0.440]	[0.801]
MOH x Cited	-0.439	-0.407	-0.462
	(0.466)	(0.511)	(0.474)
	[0.346]	[0.425]	[0.330]
Intern Org x Cited	-0.208	-0.330	0.0261
	(0.400)	(0.462)	(0.375)
	[0.604]	[0.475]	[0.945]
Observations	492	492	492
Depvar Mean	3.644	2.274	1.370

Notes: Standard errors are reported in parentheses. p-values are reported in brackets. All columns include fixed effects for phase and condition on non-exception tweets. The omitted categories are other types of sources, and its interaction with source being cited.

Appendix D. Content Heterogeneity

Table D.1. How Content Affects Retweeting by F1 likes/retweets ?

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # Likes	(4) Poisson # Pooled	(5) Poisson # Retweets
Myth-busting Facts	0.588	0.627	0.518	-0.0481	0.136
	(0.319)	(0.346)	(0.381)	(0.296)	(0.413)
	[0.0654]	[0.0698]	[0.174]	[0.871]	[0.742]
Access Info	0.402	0.319	0.530	0.315	0.477
	(0.258)	(0.292)	(0.309)	(0.294)	(0.366)
	[0.118]	[0.275]	[0.0863]	[0.284]	[0.192]
Importance Info	0.543	0.526	0.565	0.466	0.442
	(0.229)	(0.267)	(0.290)	(0.246)	(0.343)
	[0.0178]	[0.0487]	[0.0516]	[0.0578]	[0.197]
Celeb writes and tweets				1.040	1.327
				(0.283)	(0.374)
				[0.000242]	[0.000382]
Access × Celeb Direct				-0.0101	-0.299
				(0.374)	(0.451)
				[0.979]	[0.508]
Importance × Celeb Direct				0.0558	0.00945
				(0.314)	(0.406)
				[0.859]	[0.981]
Myth × Celeb Direct				0.652	0.432
				(0.381)	(0.485)
				[0.0871]	[0.374]
Access × Celeb RT Org				0.103	-0.0418
				(0.244)	(0.256)
				[0.672]	[0.870]
Importance × Celeb RT Org				0.135	0.402
				(0.215)	(0.190)
				[0.531]	[0.0348]
Myth × Celeb RT Org				0.250	0.147
				(0.290)	(0.384)
				[0.389]	[0.701]
Observations	492	492	492	492	492
Depvar Mean	3.644	3.644	3.644	3.644	3.644

Notes: Standard errors (clustered at the celebrity/organization level) are reported in parentheses. p-values are reported in brackets. All columns include fixed effects for number of non-exception tweets assigned and condition on non-exception tweets. The omitted category is non-myth facts.

Appendix E. Impact of No. of Forced Joes RTs on F2 and F1 likes/retweets

Table E.1. Discrete

VARIABLES	(1) F2 Poisson # Pooled	(2) F2 Poisson # Retweets	(3) F2 Poisson # Likes	(4) F1 Poisson # Pooled	(5) F1 Poisson # Retweets	(6) F1 Poisson # Likes
5 Forced Joe RTs assigned	0.0466	0.0399	0.125	0.365	0.444	0.221
	(0.332)	(0.346)	(1.065)	(0.371)	(0.388)	(0.388)
	[0.889]	[0.908]	[0.906]	[0.326]	[0.252]	[0.568]
10 Forced Joe RTs assigned	0.298	0.244	0.602	0.113	0.0395	0.215
	(0.417)	(0.414)	(0.852)	(0.401)	(0.440)	(0.391)
	[0.475]	[0.556]	[0.480]	[0.779]	[0.928]	[0.583]
15 Forced Joe RTs assigned	0.379	0.256	0.982	0.340	0.207	0.514
	(0.383)	(0.407)	(0.846)	(0.320)	(0.359)	(0.323)
	[0.323]	[0.529]	[0.246]	[0.288]	[0.565]	[0.112]
Observations	505	505	505	184	184	184
Phase Control	✓	✓	✓	✓	✓	✓
Log #followers control	✓	✓	✓	✓	✓	✓
Message style control	✓	✓	✓	✓	✓	✓
Depvar Mean	0.226	0.184	0.0416	4.527	2.707	1.821
1 Forced Joe RT assigned log mean	-2.197	-2.331	-4.277	1.369	0.870	0.435

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

Table E.2. Continuous

VARIABLES	(1) F2 Poisson # Pooled	(2) F2 Poisson # Retweets	(3) F2 Poisson # Likes	(4) F1 Poisson # Pooled	(5) F1 Poisson # Retweets	(6) F1 Poisson # Likes
Number of Forced Joe RTs assigned	0.0367	0.0259	0.0864	-0.000776	-0.0247	0.0340
	(0.0317)	(0.0286)	(0.0946)	(0.0434)	(0.0457)	(0.0425)
	[0.247]	[0.366]	[0.361]	[0.986]	[0.589]	[0.424]
Observations	401	401	401	140	140	140
Phase Control	✓	✓	✓	✓	✓	✓
Log #followers control	✓	✓	✓	✓	✓	✓
Message style control	✓	✓	✓	✓	✓	✓
Depvar Mean	0.244	0.197	0.0474	4.714	2.807	1.907
1 Forced Joe RT assigned log mean	-2.197	-2.331	-4.277	1.369	0.870	0.435

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

Appendix F. Effect of celebrity retweeting organizations

Table F.1. Reach vs. Endorsement: Value of Celeb Endorsement for Joes and Organizations through Involvement measured by F2 likes/retweets

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # Likes
Celeb writes and tweets	0.375	0.364	0.570
	(0.180)	(0.177)	(0.518)
	[0.0371]	[0.0392]	[0.271]
Org writes and Celeb retweets	0.597	0.632	0.300
	(0.202)	(0.224)	(0.528)
	[0.00308]	[0.00474]	[0.570]
Observations	1,899	1,899	1,899
Phase control	✓	✓	✓
Log #followers control	✓	✓	✓
Message style control	✓	✓	✓
Joe writes mean	0.0107	0.0471	0.0107

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

TABLE F.2. Value of Celeb Endorsement for Joes and Organizations through Composition measured by F1 likes/retweets

VARIABLES	(1) Poisson # Pooled	(2) Poisson # Retweets	(3) Poisson # Likes
Celeb writes and tweets	1.205	1.411	0.913
	(0.102)	(0.105)	(0.131)
	[0]	[0]	[0]
Org writes and Celeb retweets	0.0597	0.246	-0.208
	(0.135)	(0.128)	(0.180)
	[0.658]	[0.0540]	[0.247]
Observations	452	452	452
Phase control	✓	✓	✓
Log #followers control	✓	✓	✓
Message style control	✓	✓	✓
Joe 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. 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.