A Model of Information Nudges

Lucas C. Coffman
Boston College

Clayton R. Featherstone
Baylor University

Judd B. Kessler
The Wharton School, University of Pennsylvania

JULY 1, 2024

Abstract

Nudge-style interventions are popular but are often criticized for being atheoretical. We present a model of information nudges (i.e., interventions that provide useful but imperfect information about the utility of taking an action) based on Bayesian updating in a setting of binary choice. The model makes two main predictions: One, the probability of a positive treatment effect should be increasing in the baseline take-up rate. Two, across studies, as baseline rates increase from 0 to 1, the expected treatment effect has a "down–up–down" shape. A surprising corollary of both predictions is that treatment effects are expected to be *negative* for low baseline rates. We use reduced-form and structural methods to conduct a meta-analysis of 75 information nudges and corroborate both predictions. Both the meta-analysis and a novel survey of nudge experts suggest the intuition in the model is not currently known. Finally, we provide guidance for practitioners about the environments in which information nudges will positively affect a desired behavior and those in which they may backfire.

ACKNOWLEDGMENTS: We are grateful to attendees at the Stanford Institute for Theoretical Economics: Experimental Session, the Behavioral Economics Annual Meetings, Advances in Field Experiments, Methods in Experimental Economics Research, Yale Research Initiative on Innovation and Scale, The Economic Science Association North America Meetings, the Office of Evaluation Sciences, Stanford, Villanova, Wharton, Purdue, Baylor, and University of Texas at Dallas

1 Introduction

Over the past two decades, a large and growing body of empirical work has investigated the impact of "nudges" on behavior.[1] Researchers have investigated the efficacy of a number of prominent nudges across a wide variety of settings. One finding from the broad application of these interventions is that some of the most regularly effective and behaviorally intuitive nudges often fail to influence behavior (DellaVigna and Linos 2022) and sometimes influence behavior in the opposite direction than expected (i.e., they **backfire**). These failures and backfires surprise many researchers and practitioners who explicitly or implicitly assume that a nudge that works in one context will work in other, similar contexts. Such an assumption may be perfectly natural in the absence of a formal theory of how the nudge affects behavior. This paper introduces such a theory for a popular nudge: providing individuals with information about a choice they face. We call such interventions **information nudges**.

Information nudges have successfully changed behavior in a wide variety of contexts.[2] However, prominent empirical papers have found null results from providing information that might have been expected to increase a desired behavior (e.g., Allcott and Taubinsky 2015 and Avitabile and de Hoyos 2018) or have found that the nudge backfires, generating treatment effects in the opposite direction than expected, for at least some groups (e.g., Fellner, Sausgruber, and Traxler 2013, Bhargava and Manoli 2013, and Beshears, Choi, Laibson, Madrian, and Milkman 2015).

These null and negative results are deemed surprising because many practitioners only test information nudges that they view as likely to be successful based on two standard intuitions. The first intuition is that informa-

1. See Sunstein and Thaler (2008) for a detailed discussion of nudges. This work has been influential in the policy domain, spawning nudge units in the UK (called the Behavioral Insights Team), US (called the Social and Behavioral Sciences Team), and around the world. See Whitehead et al. (2014).

2. We say "successfully changed behavior" when it moves behavior in a direction hypothesized by the researcher and desired by practitioners. Information about others' decisions has affected decisions to donate money (see, e.g., Frey and Meier 2004, Martin and Randal 2008, Croson and Shang 2008, and Shang and Croson 2009), rate movies (Chen et al. 2010), order certain entrées (Cai, Chen, and Fang 2009), save energy (Allcott 2011), reuse towels (Goldstein, Cialdini, and Griskevicius 2008), pay taxes (Hallsworth et al. 2014), like certain songs (Salganik, Dodds, and Watts 2006), steal petrified wood (Cialdini et al. 2006), intend to vote (Gerber and Rogers 2009), litter (Cialdini, Reno, and Kallgren 1990), take a job (Coffman, Featherstone, and Kessler 2017), give money in a laboratory public goods games (Keser and Van Winden 2000), (Fischbacher, Gächter, and Fehr 2001), (Potters, Sefton, and Vesterlund 2005). Information about the costs or benefits of different actions has been shown to affect school choice (Hastings and Weinstein 2007), standardized test scores (Nguyen 2008), graduation rates (Jensen 2010), claiming tax benefits (Bhargava and Manoli 2013), tax compliance (Pomeranz 2015), 401(k) contribution levels (Clark, Maki, and Morrill 2014), eating fewer calories (Bollinger, Leslie, and Sorensen 2011), responding to energy price changes (Jessoe and Rapson 2014), and purchasing fluorescent light bulbs (Allcott and Taubinsky 2015).

tion nudges will be effective when the nudge is "good news" relative to the average belief in the population of agents (see, e.g., Schultz et al. 2007).[3] The second intuition is that nudges—including information nudges—are likely to be effective when there are many agents who might potentially be nudged from not taking the action to taking the action.[4]

We develop a model of information nudges to explain when they will be effective and when they will backfire. In doing so, we are able to reconcile existing results from the literature and provide guidance to practitioners of information nudges about when a treatment effect is likely to be in the desired direction and when it is likely to be large. Our model is one in which rational agents treat the information provided by a nudge as a signal that leads them to update their beliefs about the relative utility of options in a binary choice.[5] Our model shows that, under rather general assumptions, the two standard intuitions about when information nudges will be effective are wrong.

The first key insight from our model is that what matters for how an information nudge affects a binary choice is how the information contained in the nudge affects agents *at the margin*. Consequently, in contrast to the first intuition, an information nudge that is "good news" relative to the belief of the *average agent* can still backfire if the belief of the *marginal* agent (i.e., one who might change her behavior in response to being nudged) is different from the belief of the *average* agent.

The second key insight from our model is that the belief of the marginal agents is negatively correlated with the rate at which agents make choices in the absence of the nudge, which we call "baseline take-up," or simply the **baseline**, for short.[6] Formalizing these two insights, in contrast to the

3. Before writing this paper, we counted ourselves among the researchers who operated under this first intuition. In Coffman, Featherstone, and Kessler (2017), we ran a large field experiment with an information intervention and highlighted in the paper that our nudge was good news to the majority. We wrote: "note that in the control condition, the median belief is consistently 71 percent, well below 84 percent (the number provided in the treatment)."

4. When we perform a meta-analysis of information nudges that appear in the literature, we find that the median rate of take-up in the absence of the nudge is 0.34 and that roughly a third of information nudges are attempted in environments with take-up below 0.23, suggesting that practitioners explicitly or implicitly rely on this second intuition.

5. This information could be *direct information* about the costs or benefits of the outcomes (e.g., information about the returns to graduating high school) or *indirect information* that leads agents to infer that something about the costs and benefits of the actions (e.g., information that the majority of other people donate to a charity). See Vesterlund (2003) for a model of how sequential fund-raising can allow potential donors to provide information to one another about the quality of a charity. Our model is in this spirit, and inspired by this work, but considers a general information structure.

6. While fully fleshed out in Section 3, to see why there is a negative relationship between baseline take-up and the belief of the marginal agent, consider that when 90% of agents take an action, an agent at the margin (i.e., who might be induced to respond when nudged) likely has a relatively low prior (e.g., near the 10th percentile of agents' priors). Alternatively,

second intuition, the model demonstrates that a nudge will be more likely to backfire in settings when baseline take-up is low and will be more likely to generate a positive treatment effect when baseline take-up is high.

.

Combining these new insights about the direction of a treatment effect with assumptions of the density of agents' beliefs, the model generates additional predictions about when treatment effects will be smaller and larger in magnitude. In particular, the model suggests a specific ''down–up–down" shape of the relationship between baseline take-up and the sign and size of the treatment effect. As baseline take-up increases from zero to one, the treatment effect will start out at zero when the baseline is zero, become negative (*decreasing* from zero and then *increasing* back to zero), reach an intermediate zero at the baseline where the belief of the marginal agents is identical to the information provided in the nudge, and then become positive (*increasing* from zero and then *decreasing* back to zero), again reaching 0 when the baseline is one.[7]

To assess if our theoretical environment reflects contexts in which academics use information interventions, we test the predictions of our theory—about how baseline relates to the sign and magnitude of treatment effects—with a meta-analysis of 75 experiments across 22 papers that use information nudges to affect a binary outcome.[8] We present reduced-form results showing the relationship between baseline take-up and the probability of a positive treatment effect and between baseline take-up and the magnitude of the treatment effect. Our reduced-form meta-analysis finds that, even across very different experimental settings, these relationships appear as our model predicts. First, the probability of a positive treatment effect increases as the baseline rate of taking the action increases. For example, the likelihood of a negative treatment effect is 46.2 when the baseline is below 0.25 but only 12.2 when the baseline is above 0.25. Second, the observed relationship between the baseline and treatment effects across experiments in the literature suggests a down–up–down shape (though the estimates are not all significant in every test), as predicted by the model.

We additionally pioneer a new method of "structural meta-analysis" in which we assume a data generating process for experiments in the literature and then fit its key parameters (i.e., those parameters that are identified as important by the model) using the data. Our structural estimates suggest that the average information nudge in the literature falls at the 63rd per-

when 10% of agents take the action, an agent at the margin likely has a relatively high prior (e.g., near the 90th percentile of agents' priors).

7. For a graphical representation of this relationship, see Figure 1h

8. As described in Section 4, papers were collected by researchers initially and supplemented with papers provided to us in response to a request email sent to the discussion e-mail list of the Economic Science Association in December of 2015.

centile of agents' prior belief distributions and that two thirds of nudges fall between the 53rd and 72nd percentiles. These results are supportive of our initial assumption that experimenters generally pick nudges that are good news to the average agent.[9]

We also present survey evidence that awareness of the intuition within our model is low or nonexistent. The meta-analysis provides the first evidence of this as a majority of experiments in the data set are run in contexts with low baseline rates, which according to our model, is where one would expect backfires or null effects. Further, in reading the extant literature of information interventions, either those that use them or guides for running them (e.g. see Haaland, Roth, and Wohlfart 2023 for a recent, helpful, thorough guide), we never find any mention of any intuition or prescription consistent with those from our model.

We attempt to provide more direct evidence of the novelty of our model by surveying both academics and policymakers who run nudge-like interventions in the field. We find that over 80% of the sample revealed intuitions inconsistent with our theory, and no respondents explain their choices using any reasoning related to that in our model. Taken together with the meta-analysis, we show that the mechanics in our model are first order in their impact on behavior, yet the insights currently seem to be missing among experts.

Our paper makes four main contributions. The first is providing a tractable model of how populations update beliefs that in the presence of a noisy but informative signal. Our model of how an individual agent responds to a given information signal is captured by just three numbers. This simplicity is crucial when it comes to aggregating the behavior of individual into the behavior of populations. Our model threads the needle by being simple enough to be tractable, while also capturing nuance like the fact that a signal can be good news to the average member of a population and still lead to a backfire.

The second is testing a model for an important class of nudges—information nudges—that have been a focus of researchers and have been used by practitioners in myriad settings. A first wave of work on nudges focused on documenting how (sometimes large) changes in behavior can be induced by (often subtle) interventions. A second wave has deepened our understanding of nudges. Work has focused on exploring the robustness of

9. Specifically, we estimate that a nudge one standard deviation below the mean nudge is at the 53rd percentile and a nudge one standard deviation above the mean nudge is at the 72nd percentile. As is discussed in Section 4.4, an additional key parameter describing the literature is the relative size of the standard deviation of agents' prior beliefs and the standard deviation of their "thresholds," where a threshold is how high an agent's belief would have to be for her to take the action. We estimate that the standard deviation of thresholds is 3.5 times the standard deviation of prior beliefs, suggesting that agents prior beliefs are more similar to one another than their outside options, which seems quite sensible in most empirical settings.

nudges to broader settings of interest and has found that nudges assumed to work well by academics often do not work at scale (DellaVigna and Linos 2022). Other work has focused on exploring the welfare effects of common nudges (see, e.g., Carroll et al. 2009, Allcott and Kessler 2019, Bernheim, Fradkin, and Popov 2015, and Butera et al. 2022), and has emphasized that just because something impacts behavior does not necessarily imply it is welfare enhancing.[10] We aim to contribute to this second wave of work by modeling how an important class of nudges affects behavior. The key insights of our model of information nudges are straightforward ex post, but were not readily apparent ex ante (and, indeed, they seem to have been overlooked by practitioners running experiments).[11] Our model rationalizes diverse findings in the extant literature and does so while assuming rational, Bayesian agents, suggesting that information nudges need not rely on "behavioral" agents to be effective at influencing behavior and that null and negative treatment effects to information nudges can also arise absent behavioral explanations.

The third contribution is pioneering a structural approach for conducting meta-analyses. In contrast to reduced-form methods (such as analyzing data combined from multiple studies or running statistical tests with results from various studies as observations), our structural approach allows us to combine our model and the available data to estimate key parameters of the data generating process in the literature. These parameter estimates serve two purposes. First, whether or not the estimates seem sensible given our understanding of how experiments are run in practice helps to directly assess our model (e.g., we might have become skeptical of our model if we estimated that practitioners were mostly testing nudges that were bad news to the average agent). Second, the estimates are useful to understand the features of a literature and can be used as benchmarks for practitioners deciding whether to implement a particular information nudge in the future. We view our structural meta-analysis as a potentially useful tool for other researchers who have a theory driven explanation for a pattern of results in an empirical literature.

10. In our context, in which nudges provide truthful signals—and under our assumption of rational, Bayesian agents—the welfare effects of the nudges we study are unambiguously weakly positive, since they can only aid in agents making better decisions.

11. When null or negative results arose previously, many papers documenting them did not give an explanation for why the nudge did not work as expected. Those that did offer explanations generally proposed alternative stories, not based on information updating (e.g., suggesting instead the possibility of backlash in response to social information or suggesting complexity in how information was provided). The two papers that offer explanations in the direction of our theory are Fellner, Sausgruber, and Traxler (2013) and Hastings, Neilson, and Zimmerman (2015). Fellner, Sausgruber, and Traxler (2013) states: "Concerning the social information, this observation is not conclusive for a final evaluation of the treatment, as its effect is expected to depend on individuals' (heterogeneous) prior beliefs". Hastings, Neilson, and Zimmerman (2015) states: "[I]t may be the case that...the remaining students have parents who are not responsive to information on academic outcomes."

The fourth contribution is providing guidance to practitioners—including academics, firms, and policy makers—who are considering using information nudges to influence agents in practice. For these readers, we offer a number of insights in Section 6. We highlight two of the most (ex-ante) counter-intuitive insights here. First, in settings where baseline take-up is low, information nudges may backfire (particularly if they are similar to the nudges that have previously been tested in the literature in how their information content relates to agents' prior beliefs). Our estimates suggest that information nudges are likely to backfire for baseline rates below roughly 0.10, and are only quite likely to succeed for baseline rates above 0.50. Second, settings in which many agents are expected to take-up at baseline may be particularly ripe for information nudges to have big positive impacts. Given our parameter estimates, the treatment effect of the "typical" nudge is expected to have the largest positive treatment effect at a baseline of 0.75.

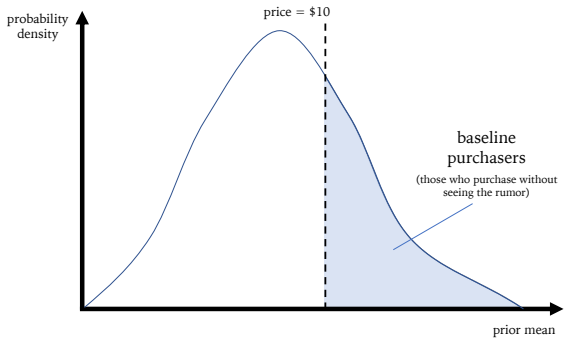
The rest of this paper proceeds as follows (and it is admittedly somewhat unusual in its structure). Since we expect and hope that some of our readers will be interested in the intuition of the model and its implications for implementing information nudges, Section 2 presents the intuition underlying information nudges without presenting much in the way of technical details. The next section, Section 3 presents a formal model. While some might skip this level of detail, we have included this theory section to provide a level of detail for theorists who are interested in how we formally model information nudges. To keep the manuscript manageable in length, many of the technical details, proofs, and extensions are omitted from the main text and instead appear in the Appendix. Section 4 presents the meta-analysis of information nudges and describes our novel method of conducting a structural meta analysis. Section 5 discusses survey measures of awareness among academics and policymakers of the intuition provided in our model. Section 6 provides guidance to practitioners. Section 7 concludes.

2 Intuition

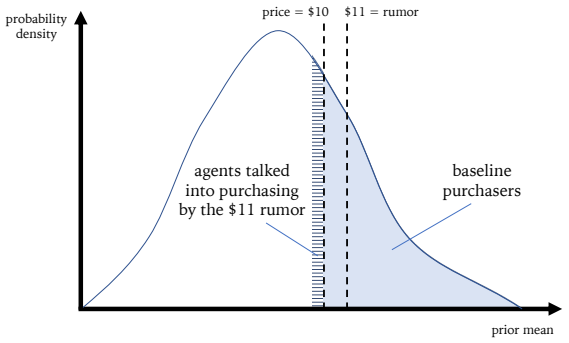
In the next section, we will introduce a formal model, but before doing so, it will prove useful to illustrate the underlying idea with an intuitive example. This is best broken into two parts. The first describes how a single experiment works, while the second describes how a literature consisting of such experiments works.

2.1 A Prototypical Information Experiment

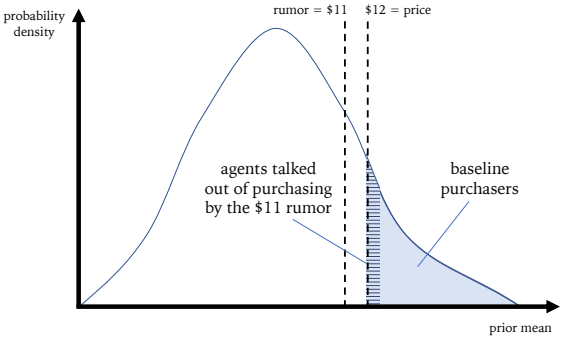
Most information experiments closely resemble to the following example. A stack of envelopes, each of which contain the same, unknown amount of cash, are being sold for $10 each. If a risk-neutral agent's prior belief



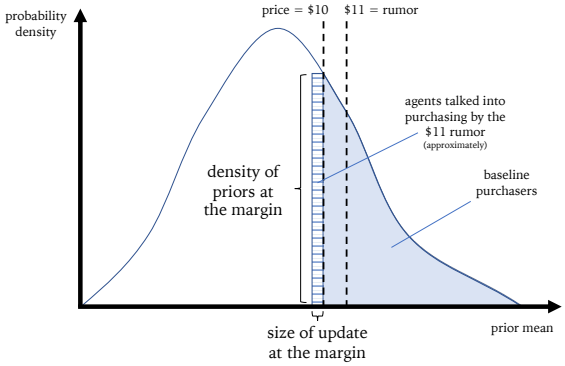
(a) When the price is $10, the baseline purchase rate is the fraction of agents whose prior exceeds $10.



(b) When the price is $10, marginal agents are *talked into* purchasing by the $11 rumor.



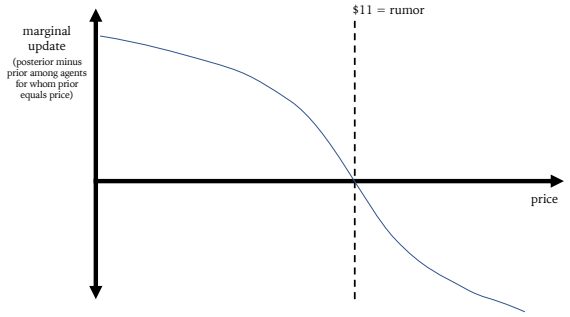
(c) When the price is $12, marginal agents are *talked out of* purchasing by the $11 rumor.



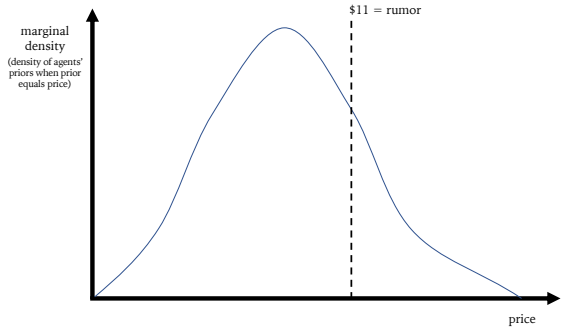
(d) When updates are small, the treatment effect—here at a $10 price—can be approximated by multiplying—among marginal agents—the density of priors and the average update.

Figure 1: The envelope example, illustrated

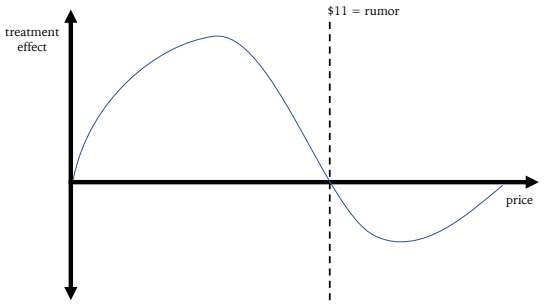
NOTES: In the subfigures above, the bell-shaped curve is the density of prior means in the population. Solid shading represents agents who purchase without being exposed to the $11 rumor, while hatching represents agents whose purchasing decisions change in response to the rumor. Vertical dashed lines represent prices or rumors.



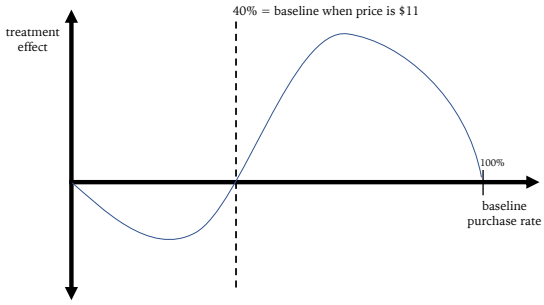
(e) An agent updates more when her her prior is further from the rumor. Hence, a marginal agent—whose whose prior is near the price—updates more when the *price* is further from the rumor.



(f) Since the price is the marginal prior, the marginal density as a function of price simply traces out the density of priors in the population.



(g) The product of parts (e) and (f) yields the treatment effect as a function of price.



(h) The law of demand dictates that baseline increases as the price decreases. Hence, treatment effect as a function of baseline is a (rescaled) mirror image of part (g).

Figure 1: The envelope example, illustrated (*continued*)

about this amount of cash has a mean greater than $10, that agent will choose to purchase; otherwise, she doesn't. In other words, the price is the **threshold** prior that separates purchasers from non-purchasers. Across the population of agents, the prior mean varies according a **belief distribution**, or equivalently, a **demand curve**.[12] Part (a) of Figure 1 shows how the **baseline** purchase rate relates to the prior distribution and the price.

Compare this baseline scenario to one in which agents are all **treated** with a rumor that the amount of cash in each envelope is actually $11. No agent puts great credence in the rumor, but each *slightly* updates her beliefs in its direction. In other words, the rumor is a **signal** being used as an **information nudge**. Small updates mean that only agents with prior means just below $10 will have their purchasing behavior changed by the rumor. The price is essentially the prior of such **marginal** agents. Hence the sign of the **treatment effect** on the purchase rate is positive because the rumor is "good news" to the marginal agent. Part (b) of Figure 1 shows this graphically.

Of course, the rumor can also be "bad news" to the marginal agent. For instance, if the price were $12 instead of $10, the set of marginal agents would be those whose beliefs are just above $12, and hence the treatment effect would be *negative*. Part (c) of Figure 1 illustrates this reversal. To be clear: the treatment effect of the *same* rumor, on the *same* population of agents, can *vary* with the price (i.e., the threshold). This is because *the price dictates the prior of the marginal agent*.

Mathematically, when updates are small, the treatment effect is approximately equal to the product—among marginal agents—of the density of priors, which is non-negative, and the average update, which can be positive or negative. This approximation, applied to part (b) of Figure 1, is illustrated in part (d). Hence, the sign of the treatment effect is determined by whether the price (i.e., the threshold)—and hence the prior of the marginal agent—is above or below the rumor (i.e., the signal).

2.2 A Literature of Information Experiments

Imagine a literature generated by versions of the envelope experiment described in the previous subsection. Different baselines and treatment effects are generated by variation in belief distribution, rumor (i.e., signal), and price (i.e., marginal prior). To develop intuition for this process, one must understand the relative importance of these sources of variation.

A key insight is that the reasoning of the previous section hinges only on where the rumor (signal) and price (marginal prior) lie *relative to the belief distribution*. Variation in the belief distribution itself isn't important. Then, it is only variation in the price and rumor (threshold/marginal prior and

12. The number of purchasers at a given price—the quantity demanded—is simply the number of agents with priors above that price.

signal) *relative to the belief distribution* that matters.

Of those two sources of variation, it makes sense to think of *price* (*marginal prior*) as the primary driver of variation in baseline and treatment effect. The reason for this is simple: **selection bias**. Those who run information-nudge experiments tend to do so in environments where their nudge is "good news" to most agents. In terms of the envelope example, this means we should not expect the rumor (signal) to vary much. Instead it will generally stay in the right shoulder of belief distribution. There is no such intuition limiting the variation in price (marginal prior).

So how are treatment effect and baseline related when their variation is driven primarily by variation in marginal prior? Imagine running the envelope experiment repeatedly with different prices (marginal priors), but holding the \$11 rumor (signal) and the belief distribution fixed. The sign of the rumor's treatment effect would be negative when the price is above the \$11 rumor and positive otherwise.

By using the demand curve, we can also phrase this regularity in terms of the baseline purchase rate. Assume that \$11 is at (for instance) the 60th percentile of the belief distribution, so that the baseline purchase rate at that price is 40%. Under this assumption, when the price (marginal prior) is above \$11, the baseline is lower than 40%; otherwise, it is higher. Hence, when the baseline is below 40%, the treatment effect is negative; otherwise, it is positive. This might seem too stark to be realistic; however, if we introduced a bit of randomness in the rumor and belief distribution, we would see a softer regularity: **the probability of a positive treatment effect would be increasing with the baseline purchase rate.**

In fact, we can push this intuition even further to find another regularity concerning, not just the sign, but the *magnitude* of the treatment effect. To do so, we need only assume that the amount by which an agent updates (i.e., the difference between her prior and posterior) is larger when the rumor is further away from her prior.[13] For example, other things equal, it is reasonable to assume an agent who believes the envelope contains \$5 will update more in response to the \$11 rumor than an agent who believes the envelope contains \$10. Since it is the price that determines who is marginal, this assumption basically says that the update of marginal agents varies with price in the way illustrated by part (e) of Figure 1. What's more, the density of priors among marginal agents will vary with price in the way illustrated by part (f) of Figure 1.

Then, looking back to the approximation mentioned at the end of the previous subsection, we see that treatment effect as a function of price (i.e., marginal prior) should look like the product of parts (e) and (f) of Figure 1, that is, like part (g). And finally, since the baseline purchase rate is mono-

13. This is the case, for instance, when the posterior is a fixed convex combination of the prior and the signal, as is true for Bayesian updating with a wide array of prior and conjugate signal distributions (Diaconis and Ylvisaker 1979).

tonically decreasing in the price (this is just the law of demand), we should expect the treatment effect as a function of the baseline purchase rate to look like part (h) of Figure 1.

That is, **the conditional-on-baseline expected treatment effect should have a "down–up–down" shape**. Intuitively, there are zeros at baselines zero and one because there are no marginal agents when literally *everyone* or *no one* purchases. The other, interior zero corresponds to the baseline that makes the prior of the marginal agent exactly equal to the signal. Such a marginal agent simply doesn't update.

For baselines *lower* than that interior zero, the treatment effect is *negative* because *low* baselines correspond to prices (marginal priors) *above* the rumor (signal), and hence marginal agents who update *down* in response. Similarly, for baselines *higher* than that interior zero, the treatment effect is *positive* because *high* baselines correspond to prices (marginal priors) *below* the rumor (signal), and hence marginal agents who update *up* in response.

To summarize, intuitively, we expect two main results. First, we expect the probability of a positive treatment effect to be increasing in the baseline. Second, we expect the conditional-on-baseline expected treatment effect to have a "down–up–down" shape like that in part (h) of Figure 1.

3 Theory

Now, we flesh out the intuition of the previous section with a formal model. We will consider three hierarchical levels: the behavior of the **agent**, the behavior of a population of agents (i.e., an **experiment**), and the behavior of a population of experiments (i.e., a **literature**). In the main text, we will discuss the simplest versions of these models, but the interested reader can find the full model with proofs in Appendix C.

3.1 Agent-Level Model

An agent must make a binary choice based on her beliefs about an unknown scalar **state**. The agent may change her choice if she is exposed to the realization of a scalar **nudge signal** that is essentially the state plus noise. We define **take-up** to be the choice encouraged by higher state values.[14]

We capture the agent's uncertainty about the state and nudge signal by treating them as the respective components of a random vector, (X, N). Unnudged, the agent makes her take-up decision based solely on her beliefs about X (i.e., with her beliefs about N integrated out). If nudged by being told the signal realization is $N = \nu$, she instead makes the take-up decision based on her updated beliefs about X, which are codified by the conditional

14. Sometimes this definition means that take-up is *not* doing something. For example, if the experimenter is trying to nudge whether the agent smokes by sending a signal about how much smoking increases cancer risk, then take-up is *not* smoking.

distribution evaluated at $N = \nu$. We assume this distribution is well defined (i.e., the support of N includes ν).

The agent is endowed with a strictly increasing and continuous **net-utility function**, u. If the agent is certain that $X = x$, the value $u(x)$ represents how much more utility she expects to get from taking up than from not taking up. To keep things on the same scale as the state, we transform expected net-utility values into **state certainty equivalents**. Let $\mu \equiv u^{-1}\big(\mathbb{E}\big[u(X)\big]\big)$ be the agent's state certainty equivalent of being un-nudged, and similarly, let $\mu' \equiv u^{-1}\big(\mathbb{E}\big[u(X)\,\big|\,N = \nu\big]\big)$ be her state certainty equivalent of being nudged. We refer to μ and μ' as the agent's **prior** and **posterior**. Note that, for the risk-neutral agent, the prior and posterior are simply the agent's expected state when un-nudged and nudged, respectively. Of course, the **nudge signal** itself, ν, can be interpreted as the state certainty equivalent of being sure that $X = \nu$.

To model the idea that the nudge is a straightforward signal about the state, we assume the agent updates towards the signal realization, but not all the way to it, so that the posterior is a convex combination of the prior and the signal.[15] Define the agent's **update strength**, ε, to be the weight on the signal in this convex combination, which, if $\nu \neq \mu$, is equal to $(\mu' - \mu)/(\nu - \mu)$.[16] By definition then, $\mu' = \mu + \varepsilon\,(\nu - \mu)$, with ε on the unit interval. In what follows, we will consider the limit as agent update strengths get small.[17]

Finally, we assume the agent takes up when her expected net-utility is weakly positive. To put this in terms of the prior and posterior, we define the agent's **threshold**, θ, to be the certain value of the state that would make her indifferent between taking up and not taking up, that is, $\theta \equiv u^{-1}(0)$.[18] Then, the un-nudged agent takes up if and only if $\mu \geq \theta$, while the nudged agent takes up if and only if $\mu + \varepsilon\,(\nu - \mu) \geq \theta$.

To summarize, an agents's potential take-up choices in a given experiment can be characterized by four numbers: her prior, μ; her threshold, θ; her update strength, ε; and the signal realization, ν. Making the connection to the envelope example, the prior is the expected amount of money in the envelopes without hearing the rumor, the signal realization is the rumor, and the threshold is the price. The update strength can be combined with the first two of these to yield the posterior, which is the expected amount

15. To be clear, we don't require the posterior to be the *same* convex combination of prior and signal, regardless of the signal, like in the paper mentioned in Footnote 13. Our assumption is essentially the more general *updating towards the signal* concept of Chambers and Healy (2012).

16. If an agent has $\mu = \nu$, any update strength works; we choose $\varepsilon = 0$. Ultimately, this won't matter, as we will assume the set of such agents is measure zero in the population.

17. To be clear, we don't require *all* agents to have small update strengths, we simply assume they are small *in expectation* across the population. For a much more precise treatment of this idea, see Section C.2 of the appendix.

18. We assume θ exists, so that the agent's take-up decision is nontrivial.

of money in the envelopes once the rumor is heard.

> **In the Appendix.** In the main text, we have confined ourselves to a model where the agent only cares about the realization of one scalar variable, X. Section C.1 shows how a richer model with utility-relevant random variables in addition to X can still be reduced to the model discussed above.

3.2 Experiment-Level Model

An **experiment** is a population of agents being potentially exposed to a common information nudge, v. To model this, we consider a random vector, (Θ, M, E), that represents the respective threshold, prior, and update strength of an agent that is uniformly and randomly selected from the population.

The fraction of agents who take-up without being nudged—i.e., the **baseline**, β—is defined by

$$\beta \equiv \Pr\{\Theta \leq M\},$$

while the fraction who take-up when nudged equals $\Pr\{\Theta \leq M + E(v - M)\}$. (The inequalities are weak because we assume the agent takes-up when her belief equals her threshold.) We define the **exact treatment effect**, τ^e, to be the change in the take-up rate caused by the nudge, that is

$$\tau^e \equiv \Pr\{M < \Theta \leq M + E(v - M)\} - \Pr\{M + E(v - M) < \Theta \leq M\}. \tag{1}$$

The two probabilities represent agents who are nudged *into* and *out of* take-up, respectively. Note that the realization of the signal, v, must *exceed* the realization of the prior, M, for those nudged into take-up and must *be exceeded by* the realization of M for those nudged out of take-up.

Looking at Equation 1, if the update strength, E, is usually small, then among those whose take-up decision is affected by the nudge, the prior and threshold, M and Θ, are usually close to each other. It will prove useful to think in terms of this closeness. Define the **deficit**, $\Delta \equiv \Theta - M$, to be the amount by which an un-nudged agent's prior would need to change to make her indifferent about whether to take-up. The nudge signal changes an agent's take-up decision if and only if Δ is between 0 and $E(v - M)$.[19] Hence, the treatment effect depends mainly on agents with deficits near zero. We call such agents **marginal**.

In terms of this change of variables, we can equivalently write our model of the experiment in terms of the random vector (Δ, M, E). In this transformed model, the baseline is given by $\beta = \Pr\{\Delta \leq 0\}$, and the treatment

19. More precisely, Δ must be on one of the two ranges in Equation 2. Which one depends on whether v exceeds or is exceeded by M.

effect is given by

$$\tau^e \;=\; \Pr\{0 < \Delta \le E(v-M)\} \;-\; \Pr\{E(v-M) < \Delta \le 0\}. \tag{2}$$

Since these probabilities are generated by narrow deficit ranges, we can approximate them with the product of the probability density of marginal deficits and the width of the ranges.

To operationalize this idea, we begin by using the law of iterated expectations to re-write Equation 2 as

$$\tau^e \;=\; \mathbb{E}\left[\; \Pr\left\{ 0 < \Delta \le E(v-M) \;\middle|\; M, E \right\} \;-\; \Pr\left\{ E(v-M) < \Delta \le 0 \;\middle|\; M, E \right\} \;\right],$$

where the probabilities are captured by the conditional-on-(M,E) distribution of Δ, whose density we will write as $f^{\Delta|ME}$. We can approximate both conditional probabilities with a single expression: $f^{\Delta|ME}(0\,|\,M,E)\;E(v-M)$. Doing so leads us to our **approximate treatment effect**, τ, given by

$$\tau \;\equiv\; \mathbb{E}\left[\, f^{\Delta|ME}(0\,|\,M,E)\;E(v-M) \,\right]. \tag{3}$$

The density in the expectation embodies the intuition that only marginal agents can contribute, since it is only they that can be persuaded to change their take-up decisions by a weak intervention (i.e., a nudge). We can sharpen this intuition with a bit of algebra by rewriting the previous expectation as

$$\tau \;=\; f^{\Delta}(0)\;\mathbb{E}\left[\, E(v-M) \;\middle|\; \Delta = 0 \,\right],$$

where $f^{\Delta}(0) \equiv \mathbb{E}\left[f^{\Delta|ME}(0\,|\,M,E) \right]$ is the unconditional deficit density among marginal agents.[20] We can further decompose the factors that drive the

20. In Section C.2, we show the general equivalence of the two expressions for τ mentioned above. Here, we limit ourselves to showing the equivalence when the random vector (Δ, M, E) has a joint density, $f^{\Delta ME}$, whose marginals are always non-zero. Then,

$$f^{\Delta|ME}\left(0\,\middle|\,\mu,\varepsilon\right) = \frac{f^{\Delta ME}(0,\mu,\varepsilon)}{f^{ME}(\mu,\varepsilon)}, \quad \text{and}$$
$$f^{ME|\Delta}\left(\mu,\varepsilon\,\middle|\,0\right) = \frac{f^{\Delta ME}(0,\mu,\varepsilon)}{f^{\Delta}(0)}.$$

where $f^{ME}(\mu,\varepsilon) \equiv \int f^{\Delta ME}(\delta,\mu,\varepsilon)\,d\delta$. These expressions mean that the original expectation we used to define the approximate treatment effect can be written as the integral

$$\tau = \iint \varepsilon(v-\mu)\,\frac{f^{\Delta ME}(0,\mu,\varepsilon)}{f^{ME}(\mu,\varepsilon)}\,f^{ME}(\mu,\varepsilon)\,d\mu\,d\varepsilon,$$

while the second expectation can be written as the integral

$$\tau = f^{\Delta}(0)\iint \varepsilon(v-\mu)\,\frac{f^{\Delta ME}(0,\mu,\varepsilon)}{f^{\Delta}(0)}\,d\mu\,d\varepsilon.$$

The equivalence of the two expressions for τ follows immediately.

expectation by writing [21]

$$\tau = f^{\Delta}(0)\, \mathbb{E}\left[E \,\middle|\, \Delta = 0 \right] \left\{ \nu - \frac{\mathbb{E}\left[E\, M \,\middle|\, \Delta = 0 \right]}{\mathbb{E}\left[E \,\middle|\, \Delta = 0 \right]} \right\}. \tag{4}$$

Intuitively then, τ is driven by three forces. First, it is driven by how many marginal agents there are, captured by the density $f^{\Delta}(0)$. More marginal agents means larger treatment effects. Second, it is driven by how strongly the marginal agents update, captured by $\mathbb{E}\left[E \middle| \Delta = 0\right]$. Nudges that—*ceteris paribus*—cause larger updates drive larger treatment effects. And third, it is driven by the difference between the signal, ν, and the quantity

$$\frac{\mathbb{E}\left[E\, M \,\middle|\, \Delta = 0 \right]}{\mathbb{E}\left[E \,\middle|\, \Delta = 0 \right]}. \tag{5}$$

When priors and update strengths are statistically independent, this is just the expected prior among marginal agents, $\mathbb{E}\left[M \middle| \Delta = 0\right]$. When they aren't, Expression 5 can be thought of as an expected prior among marginal agents where agents with larger update strengths are more heavily weighted.[22]

Note that this third factor is the only one that can be negative; hence, any backfire (as described in the introduction) must be driven by a nudge signal that is, on average, bad news *to marginal agents*. Here, the math highlights an important subtlety: on average, a nudge could simultaneously be bad news to marginal agents while being good news to the population as a whole. For a binary choice though, it's the marginal agents that matter.

> **In the Appendix.** In the main text, we confine ourselves to deriving our approximation quite heuristically. In Section C.2 we formally bound the error introduced by approximating τ^e with τ. This bound only requires that the density $f^{\Delta|ME}$ exist for agents whose deficits are in some neighborhood of zero. It is also robust to "poorly behaved" densities that have discontinuities, that asymptote to infinity, or that have slopes that asymptote to infinity. Such robustness is crucial when modeling social-information interventions that have priors, posteriors, and thresholds that share the unit interval as their outcome space. On the unit interval, even common densities, like that of the beta distribution, are often "poorly behaved."

21. Tacitly, we are assuming that $\mathbb{E}\left[E \middle| \Delta = 0\right] > 0$. To assume otherwise would be to assume that marginal agents ignore the nudge, a case which our model is not designed to handle.

22. Using the law of iterated expectations, the ratio in Expression 5 can be written

$$\mathbb{E}\left[\frac{E}{\mathbb{E}\left[E \middle| \Delta = 0\right]}\, \mathbb{E}\left[M \middle| \Delta = 0, E\right] \,\middle|\, \Delta = 0 \right].$$

Intuitively, among marginal agents, we look at the mean prior for each constant update-strength population and then average across populations, placing greater weight on those with larger update strengths. Among marginal agents, the expectation of the weights is 1.

In addition, we formalize the idea that the bound mentioned in the previous paragraph is tight by considering an asymptotic sequence of experiments in which the expected update strength converges to zero. Given such a sequence, the approximate treatment effect, τ, is **asymptotically equivalent** to the exact treatment effect, τ^e, if the ratio τ^e/τ converges to one. We show this to be true under two conditions.

First, the ratio $\left|\mathbb{E}\left[E(v-M)\,\middle|\,\Delta=0\right]/\mathbb{E}\left[E(v-M)\right]\right|$ must be bounded. This ensures that the updates of the the marginal agents aren't too different from those in the general population, even asymptotically. Second, the ratio $\mathbb{E}\left[E^2(v-M)^2\right]/\mathbb{E}\left[E(v-M)\right]$ must converge to zero. Intuitively, this says that the error from a first-order Taylor approximation in $E(v-M)$ is asymptotically dominated by the approximation itself. While it might seem like this is simply true, it must be assumed. The assumption rules out having the treatment effect be largely driven by agents with large updates, even though the expected update approaches zero.[23]

3.3 Literature-Level Model

In the previous two subsections, we introduced the agent- and experiment-level models. They were intended to hold quite generally, with relatively few assumptions. Our literature-level model must necessarily deviate from such an approach. Why? Looking back to the literature-level intuition discussion of Section 2, our results followed from the assumption that variation in thresholds is the primary driver of variation in baseline and treatment effect across the literature. Formalizing this logic will obviously require making assumptions that could potentially be incorrect. Of course, whether this is the case is ultimately an empirical question—one we will address later on in the paper.

3.3.a Modeling variation across the literature

The big idea of our literature-level model is that variation in the baseline and treatment effect is driven by variation in the nudge signal and the distribution of (Θ, M, E). To capture the first idea, at the literature level, we model the nudge signal as the realization of a random variable, N. To capture the second idea, we assume the distribution of (Θ, M, E) is changed by some random vector of **literature-level parameters**, Π. Concatenating our sources of variation, we are essentially assuming that the literature is generated by drawing experiments as realizations of the random vector (N, Π).

Before elaborating on this idea, however, it will first prove helpful to introduce two modifications to the experiment-level model. The first aids comparison across experiments which could ostensibly be from very different contexts; the second is merely a simplifying assumption that makes

23. For instance, consider a situation where $E(v-M)$ equals 1 with probability α and 0 otherwise, and let α approach 0. Then, $\mathbb{E}\left[E(v-M)\right] = \alpha$, but so does $\mathbb{E}\left[E^2(v-M)^2\right]$. It makes sense that our approximation wouldn't work in this situation as the treatment effect is driven entirely by agents with large updates—those with update 0 don't contribute at all.

what follows tractable.

3.3.b Simplifying the experiment-level model

Across a literature, there are many experiments in many different settings. Ultimately, we will need such different experiments to be directly comparable. How can we do this? Recall Section 2, where we discussed how, intuitively, what drives variation in a literature is the variation of thresholds and the signal *relative to the prior distribution*. To take advantage of this intuition then, we will transform our model so that the prior distribution in all experiments is the same. Doing so unifies the interpretation of priors, posteriors, thresholds, and nudge signals.

To accomplish this goal, we begin by considering what would happen if we took an experiment and applied the same strictly increasing function—call it H—to μ, μ', ν, and θ. Then, $H(\mu')$ would still be a convex combination of $H(\nu)$ and $H(\mu)$,[24] the un-nudged agent would take up if and only if $H(\mu) \geq H(\theta)$, and the nudged agent would take up if and only if $H(\mu') \geq H(\theta)$. Hence, if we defined a new prior, posterior, and threshold by $H(\mu)$, $H(\mu')$, and $H(\theta)$, we would have new version of the Section 3.1 model that makes the same behavioral predictions. In other words, **the agent-level model is invariant under strictly increasing transformations.**

Using the freedom this invariance grants us, we can transform any experiment to have any distribution of priors by simply mapping quantiles of the original prior distribution to quantiles of the desired reference distribution. Going forward then, **we will assume that the prior distribution in all experiments is the standard normal.**[25] Given this change, the μ, μ', ν, and θ values for a given agent in a given experiment should now be interpreted as z-scores against the original prior distribution from that experiment.

So, the first change to the experiment-level model is essentially without loss of generality: by thinking in terms of prior-distribution z-scores, we make it possible to compare more readily across experiments. Our second change is more substantive. Intuitively, thresholds summarize preferences, while the prior and update strength summarize information processing. Going forward, **we will assume, for any given experiment, that thresholds are independent of priors and update strengths.**

Given the two, just-discussed modifications to the experiment-level model, for a given experiment, we can now write the density of the (Θ, M, E) vector, $f^{\Theta ME}$, in a particular way: as the product of a threshold density, f^{Θ}, a conditional-on-prior update-strength density, $f^{E|M}$, and the prior

24. And hence, there would be a new update strength on the unit interval, $\tilde{\varepsilon}$ such that $H(\mu') = H(\mu) + \tilde{\varepsilon}\big(H(\nu) - H(\mu)\big)$.

25. If F^M is the original prior distribution in the experiment, this is accomplished by setting the H of the previous paragraph to $\Phi^{-1} \circ F^M$, where Φ^{-1} is the inverse distribution function of the standard normal.

density, which is always that of the unit normal, φ. In other words, we are assuming that $f^{\Theta ME}(\theta, \mu, \varepsilon) = f^{\Theta}(\theta)\ f^{E|M}(\varepsilon \,|\, \mu)\ \varphi(\mu)$.

3.3.c The literature's data-generating process

We are modeling a randomly drawn experiment in the literature as a realization of the random vector (N, Π). The realization of N is the nudge signal for the experiment, while the realization of Π moves the densities underlying (Θ, M, E). We denote this with subscripts: when the realization of Π is π, the densities that describe (Θ, M, E) are f^{Θ}_{π} and $f^{E|M}_{\pi}$.[26]

Without loss of generality, we assume one of the literature-level parameters is B, the baseline of the experiment represented by Π.[27] We then denote the rest of the literature-level parameters by Λ, calling them the **literature-level noise**. So, the random vector Π can be written as the concatenation of the random variable, B, which represents the baseline of the drawn experiment, and the random vector Λ, that is, $\Pi = (B, \Lambda)$.

In Section C.3 of the theory appendix, we will consider the model with literature-level noise, but for the sake of simplicity, we will ignore it in the main text, so that B is the only parameter driving variation in the distribution of (Θ, M, E). Going forward then, we will consider a literature modeled by the random vector (N, B). A realization, (ν, β) of (N, B), represents an experiment whose nudge signal is ν and whose (Θ, M, E) vector is described by the density $f^{\Theta ME}_{\beta}(\theta, \mu, \varepsilon) = f^{\Theta}_{\beta}(\theta)\ f^{E|M}_{\beta}(\varepsilon \,|\, \mu)\ \varphi(\mu)$.

3.3.d Baseline

If the realization of B is β, the baseline of the associated experiment must be β, that is, the equation

$$\beta = \int_{-\infty}^{\infty} F^{\Theta}_{\beta}(\mu)\, \varphi(\mu)\, d\mu \tag{6}$$

must hold (where $F^{\Theta}_{\beta}(\theta) \equiv \int_{-\infty}^{\theta} f^{\Theta}_{\beta}(\tilde{\theta})\, d\tilde{\theta}$ is the distribution associated with the density f^{Θ}_{β}).[28]

Looking at Equation 6, two things become clear. First, an increase in baseline must somehow increase the distribution function F^{Θ}_{β}. (Recall that *increasing* this distribution function makes *lower* values of Θ more likely.) A shift down in the likelihood-ratio sense is a natural way to model this. Going forward, **we will assume that increasing β shifts the distribution of Θ down in the likelihood-ratio sense.** Formally this means that, for

26. Per the previous subsection, regardless of π, the prior density is that of the standard normal, φ.

27. It could be that, while B is one of the components of Π, the distribution of (Θ, M, E) is independent of that particular component.

28. To see that the integral in Equation 6 must represent the baseline, note that the mass of agents with priors between μ and $\mu + d\mu$ is $\varphi(\mu)\, d\mu$, and that of those agents, the fraction whose threshold is less than μ is $F^{\Theta}_{\beta}(\mu)$.

any two baselines, β and β', and any two thresholds, θ and θ', if $\beta' > \beta$ and $\theta' > \theta$, then

$$\frac{f^{\Theta}_{\beta'}(\theta')}{f^{\Theta}_{\beta'}(\theta)} < \frac{f^{\Theta}_{\beta}(\theta')}{f^{\Theta}_{\beta}(\theta)}.$$

Since the likelihood-ratio reflects the likelihood of a higher threshold relative to that of a lower one, this inequality simply says that higher thresholds are relatively more likely at lower baselines. To see how this feeds back to Equation 6, recall the well-known fact that if increasing β shifts Θ down in the likelihood-ratio sense, then it also shifts Θ down in the first-order stochastic sense, so that F^{Θ}_{β} is *increasing* in β. Then, the integrand in Equation 6 increases when β increases, and our decreasing-likelihood-ratio assumption is indeed internally consistent.

The other thing that becomes clear upon looking at Equation 6 is that the expected threshold approaches $+\infty$ as the baseline approaches zero. To see this, note that for the baseline to approach zero, the distribution $F^{\Theta}_{\beta}(\mu)$ must approach zero almost everywhere. But since $F^{\Theta}_{\beta}(\mu)$ must also go from zero to one as μ goes from $-\infty$ to $+\infty$, it must be that the mass in the threshold distribution gets pushed to larger and larger values (where the density φ is small) as the baseline approaches zero. A similar argument establishes that the expected threshold approaches $-\infty$ as the baseline approaches one. **To summarize, as the baseline goes from zero to one, the expected threshold goes from positive to negative infinity.**[29]

3.3.e The conditional-on-baseline probability of a positive treatment effect

Looking to Equation 4, we see the conditional-on-baseline treatment effect is positive if the realization, ν, of the signal, N, exceeds

$$\frac{\mathbb{E}_{\beta}\left[\, E\, M \,\middle|\, \Delta = 0 \,\right]}{\mathbb{E}_{\beta}\left[\, E \,\middle|\, \Delta = 0 \,\right]}, \tag{7}$$

where the expectations have β subscripts to remind us of that dependence. Hence, if we let $G^{N|B}$ represent the conditional-on-baseline distribution of N, we can write the conditional-on-baseline probability of a positive treatment effect as

$$\Pr\{\, \tau_{B,N} > 0 \;\mid\; B = \beta \,\} \;=\; 1 - G^{N|B}\left(\frac{\mathbb{E}_{\beta}\left[E\, M \middle| \Delta = 0 \right]}{\mathbb{E}_{\beta}\left[E \middle| \Delta = 0 \right]} \;\middle|\; \beta \right), \tag{8}$$

where the (B, N) subscripts on τ remind us of that dependence. Whether this is increasing in baseline depends on how Expression 7 and $G^{N|B}\left(\nu \middle| \beta\right)$ change with β. We will consider these two comparative statics in turn. To

29. Making this argument more rigorous involves concepts like vague convergence, the monotone convergence theorem, and the integrated tail-probability expectation formula (Lo 2019). We defer these more esoteric details to the theory appendix.

address the first, it will prove helpful to write Expression 7 in a particular form.

Result. *Expression 7 can be written*

$$\frac{\mathbb{E}_\beta\left[\, E\, M \,\middle|\, \Delta = 0 \,\right]}{\mathbb{E}_\beta\left[\, E \,\middle|\, \Delta = 0 \,\right]} = \int_{-\infty}^{\infty} \mu\, \psi_\beta(\mu)\, d\mu,$$

where ψ_β is a density defined by

$$\psi_\beta(\mu) \equiv \frac{\mathbb{E}_\beta\left[E \,\middle|\, M = \mu\right] f_\beta^\Theta(\mu)\, \varphi(\mu)}{\int_{-\infty}^{\infty} \mathbb{E}_\beta\left[E \,\middle|\, M = \tilde{\mu}\right] f_\beta^\Theta(\tilde{\mu})\, \varphi(\tilde{\mu})\, d\tilde{\mu}}.$$

Proof. We will begin by deriving the conditional-on-Δ density of (M, E), as the expectations in Expression 7 will require it. Note that because $\Delta = \Theta - M$, the density of Δ conditional on $(M, E) = (\mu, \varepsilon)$ is just the threshold density shifted over by μ, that is, $f_\beta^{\Delta|ME}(\delta \,|\, \mu, \varepsilon) = f_\beta^\Theta(\delta + \mu)$. Then by the definition of conditional densities, the joint density of (Δ, M, E) must be $f_\beta^{\Delta ME}(\delta, \mu, \varepsilon) = f_\beta^{\Delta|ME}(\delta \,|\, \mu, \varepsilon)\, f_\beta^{ME}(\mu, \varepsilon)$, which is equal to $f_\beta^\Theta(\delta + \mu)\, f_\beta^{ME}(\mu, \varepsilon)$, which is equal to $f_\beta^\Theta(\delta + \mu)\, f_\beta^{E|M}(\varepsilon \,|\, \mu)\, \varphi(\mu)$. Finally then, we see that the conditional-on-Δ density of (M, E) can be written $f_\beta^{ME|\Delta}(\mu, \varepsilon \,|\, \delta) = f_\beta^{\Delta ME}(\delta, \mu, \varepsilon) / f_\beta^\Delta(\delta)$, where f_β^Δ is the marginal density of Δ. Hence, we can compute $f_\beta^{ME|\Delta}(\mu, \varepsilon \,|\, 0)$ as $f_\beta^{\Delta ME}(0, \mu, \varepsilon) / f_\beta^\Delta(0)$, which yields

$$f_\beta^{ME|\Delta}(\mu, \varepsilon \,|\, 0) = \frac{f_\beta^\Theta(\mu)\, f_\beta^{E|M}(\varepsilon \,|\, \mu)\, \varphi(\mu)}{f_\beta^\Delta(0)}.$$

Fortunately, $f_\beta^\Delta(0)$ will cancel out of Expression 7, so there is no need to explore it further.

Using the derivation of the previous paragraph, in terms of our literature-level model, we can write Expression 7 as

$$\frac{\mathbb{E}_\beta\left[\, E\, M \,\middle|\, \Delta = 0 \,\right]}{\mathbb{E}_\beta\left[\, E \,\middle|\, \Delta = 0 \,\right]} = \frac{\int_{-\infty}^{\infty} \int_0^1 \mu\, \varepsilon\, f_\beta^\Theta(\mu)\, f_\beta^{E|M}(\varepsilon \,|\, \mu)\, \varphi(\mu)\, d\varepsilon\, d\mu}{\int_{-\infty}^{\infty} \int_0^1 \varepsilon\, f_\beta^\Theta(\mu)\, f_\beta^{E|M}(\varepsilon \,|\, \mu)\, \varphi(\mu)\, d\varepsilon\, d\mu},$$

We can then simplify by writing the inner integrals as conditional-on-prior expectations to get

$$\frac{\mathbb{E}_\beta\left[\, E\, M \,\middle|\, \Delta = 0 \,\right]}{\mathbb{E}_\beta\left[\, E \,\middle|\, \Delta = 0 \,\right]} = \frac{\int_{-\infty}^{\infty} \mu\, \mathbb{E}_\beta\left[E \,\middle|\, M = \mu\right] f_\beta^\Theta(\mu)\, \varphi(\mu)\, d\mu}{\int_{-\infty}^{\infty} \mathbb{E}_\beta\left[E \,\middle|\, M = \mu\right] f_\beta^\Theta(\mu)\, \varphi(\mu)\, d\mu}. \tag{9}$$

From here, it is clear that

$$\frac{\mathbb{E}_\beta\left[\, E\, M \,\middle|\, \Delta = 0 \,\right]}{\mathbb{E}_\beta\left[\, E \,\middle|\, \Delta = 0 \,\right]} = \int_{-\infty}^{\infty} \mu\, \psi_\beta(\mu)\, d\mu,$$

where ψ_β is defined in the statement of the result. And ψ_β is indeed a density, as it is never negative and integrates out to one. □

So, why have we derived this expression? Note the likelihood ratio for ψ_β is

$$\frac{\psi_\beta(\mu')}{\psi_\beta(\mu)} = \frac{\mathbb{E}_\beta\left[E \middle| M=\mu'\right]}{\mathbb{E}_\beta\left[E \middle| M=\mu\right]} \frac{f^\Theta_\beta(\mu')}{f^\Theta_\beta(\mu)} \frac{\varphi(\mu')}{\varphi(\mu)}.$$

The likelihood-ratio assumption from the previous subsection states that the middle ratio on the right-hand side is decreasing in β. Clearly, the rightmost ratio doesn't change with β. Hence, the entire likelihood ratio for ψ_β is decreasing in β so long as $\mathbb{E}_\beta\left[E \middle| M=\mu'\right] / \mathbb{E}_\beta\left[E \middle| M=\mu\right]$ doesn't increase too much with β. Intuitively, this simply says that increasing the baseline doesn't increase the update strengths of those with high priors too much relative to those with low priors.

Going forward, we will assume this is the case: formally, for any two baselines, β and β', and any two thresholds, θ and θ', if $\beta' > \beta$ and $\theta' > \theta$, then

$$\frac{\psi^\Theta_{\beta'}(\theta')}{\psi^\Theta_{\beta'}(\theta)} < \frac{\psi^\Theta_\beta(\theta')}{\psi^\Theta_\beta(\theta)}.$$

Given this assumption, we then know that $\mathbb{E}_\beta\left[\, E\, M \,\middle|\, \Delta = 0\,\right] / \mathbb{E}_\beta\left[\, E \,\middle|\, \Delta = 0\,\right]$ is decreasing in β.

Now, looking back to Equation 8, we see that the whole thing must be increasing in β if $G^{N|B}$ is decreasing in baseline. Intuitively, this assumption means that increasing β weakly increases the signal distribution in the first-order stochastic sense.

> **Prediction 1.** If increasing β weakly increases the signal distribution in the first-order stochastic sense (i.e., if $\partial G^{N|B} / \partial \beta \leq 0$), then the conditional-on-baseline probability of a positive treatment effect is increasing in the baseline.

Of course, the assumption in Result 1 is stronger than is necessary. So long as increasing the baseline doesn't shift down the distribution of N too much, the conditional-on-baseline probability of a positive treatment effect will still be increasing in baseline.

Before moving on, it is worth heuristically describing the conditions that give rise to Prediction 1. Essentially, we needed that $\mathbb{E}_\beta\left[E \middle| M=\mu'\right] / \mathbb{E}_\beta\left[E \middle| M=\mu\right]$ and $G^{N|B}$ not increase too much with baseline. Intuitively, the first condition prevents increasing the baseline from increasing too much the update strengths (and hence the contribution to the treatment effect) of agents with high priors. Such agents are those who view the signal as bad news about take-up. The second condition prevents increasing baseline from pushing the signal down too much. Low signals are obviously more likely to be perceived as bad news.

3.3.f Conditional-on-baseline expected treatment effect

In terms of our literature-level model, we can take the conditional-on-baseline expectation of Equation 3 and write it as

$$\mathbb{E}\left[\tau_{B,N} \middle| B = \beta\right] = \int_{-\infty}^{\infty} \int_{0}^{1} \left(\mathbb{E}\left[N \middle| B = \beta\right] - \mu\right) \varepsilon \, f_{\beta}^{\Theta}(\mu) \, f_{\beta}^{E|M}\left(\varepsilon \middle| \mu\right) \varphi(\mu) \, d\varepsilon \, d\mu.$$

Through similar methods to those used in the previous subsection, we can then write this in terms of ψ_β as

$$\mathbb{E}\left[\tau_{B,N} \middle| B = \beta\right] = \left\{\int_{-\infty}^{\infty} \mathbb{E}_{\beta}\left[E \middle| M = \mu\right] f_{\beta}^{\Theta}(\mu) \, \varphi(\mu) \, d\mu \right\} \times \left\{\int_{-\infty}^{\infty} \left(\mathbb{E}\left[N \middle| B = \beta\right] - \mu\right) \psi_{\beta}(\mu) \, d\mu\right\}. \quad (10)$$

We now analyze the two terms of right-hand side of the equation in turn.

The first is clearly positive, and we expect it to be larger where there is more overlap between the prior and threshold distributions, since the integrand is determined by the product of those densities. Intuitively then, our term should approach zero as the baseline approaches either zero or one, since extreme baselines will shift thresholds to very high and very low values, where the prior distribution, φ, is quite thin. (For a more in depth discussion of why this is so, see Section 3.3.d above.) **So, the first term is positive, but approaches zero when the baseline approaches either zero or one.**

Moving on to the second term, it is helpful to consider changing the β in the expectation and the β in the subscript independently. Increasing the β in the subscript will clearly increase the second term, since $\mathbb{E}\left[N \middle| B = b\right] - \mu$ is decreasing in μ and we are assuming that the density $\psi_\beta(\mu)$ is decreasing in the likelihood-ratio sense. Increasing the β in the expectation will also increase the second term if we assume that $\mathbb{E}\left[N \middle| B = b\right]$ is weakly increasing in b.

So, putting together these two ideas, we find that **the second term in Equation 10 is weakly increasing in $\boldsymbol{\beta}$ if $\mathbb{E}\left[\boldsymbol{N} \middle| \boldsymbol{B} = \boldsymbol{\beta}\right]$ is increasing in β.** (As before, this last assumption is stronger than is needed; really, we just need that $\mathbb{E}\left[N \middle| B = \beta\right]$ doesn't decrease too much in β.) In addition, **the second term gets negative as the baseline approaches zero and positive as the baseline approaches one**, for reasons that parallel the discussion in Section 3.3.d above.[30]

Putting together the bolded points above, we find that the conditional-on-baseline expected treatment effect has a specific "down–up–down" shape to it. Formally,

30. See footnote 29.

Prediction 2 (down–up–down). As baseline approaches zero or one, $\mathbb{E}\left[\tau_{B,N} \middle| B = \beta\right]$ approaches zero. $\mathbb{E}\left[\tau_{B,N} \middle| B = \beta\right]$ also has an interior zero at some baseline strictly greater than zero and strictly less than one. Below this interior zero, $\mathbb{E}\left[\tau_{B,N} \middle| B = \beta\right]$ is negative; above it $\mathbb{E}\left[\tau_{B,N} \middle| B = \beta\right]$ is positive.

Note that, while this result lines up well with the shape in part (h) of Figure 1, the single-peakedness of the negative and positive parts of the curve is not required. What is required is that as the baseline increases from zero to one, the conditional-on-baseline expected treatment effect must 1) start at zero, 2) decrease to negative values for lower baselines, 3) come back to zero, 4) go up to positive values for higher baselines, and then 5) descend back to zero as the baseline approaches one.

Before moving on, it is worth heuristically describing the conditions that give rise to Prediction 2. Essentially, we needed that $\mathbb{E}_\beta\left[E \middle| M = \mu'\right] / \mathbb{E}_\beta\left[E \middle| M = \mu\right]$ not increase too much with baseline and $\mathbb{E}\left[N \middle| B = \beta\right]$ not decrease too much. The intuition underlying these conditions is essentially identical to that described at the end of the previous subsection.

In the Appendix. In the main text, we simplify as much as possible and keep most discussions to the heuristic. As mentioned at the beginning of this section, a more formal treatment with proofs can be found in Section C.3 of the appendix.

In addition, the appendix holds two extra classes of more technical results for the interested reader. First, the treatment there allows for our so-called literature-level noise, which is effectively variation in the distribution of (Θ, M, E) that doesn't affect the baseline. This is a significant generalization to the model presented in the main text. Second, we deal much more rigorously with the limits as baseline approaches zero and one. Essentially, the difficulty is that for baselines to really get large or small, threshold distributions must get very large or very small. Dealing with the problem of probability mass escaping at infinity is a significant technical challenge. (See footnote 29.)

4 Meta-Analysis

In this section, we test if the environment constructed in the theory is reflective of contexts where information interventions are common practice. Specifically, we simultaneously test two empirical questions. First, are the assumptions in the model (e.g., in Section 3.3 the model assumes, "that thresholds are independent of priors and update strengths") largely true for the environments in previous information experiments? Second, is the impact on behavior from the mechanisms in our theory meaningful in magnitude? Behavior in and across experiments may vary for hundreds of reasons.

Is the specific impact from our theory of first order importance alongside all of these other factors?

To test these questions, we will see if predictions from our theory hold using a meta-analysis of 75 experiments across 22 papers that use information nudges to affect a binary outcome in the extant literature. If we find evidence to support our predictions across these experiments, then we will conclude that both our assumptions largely hold in these studies, and that our mechanism's impact on behavior is important.

As described in Section 2 and Section 3, the model makes two predictions about the relationship between baseline take-up and the treatment effect. The first prediction (Result 1) is that the likelihood of a negative treatment effect will be larger at lower baselines and smaller at higher baselines. The second prediction (Result 2) is that the treatment effect will follow the down–up–down pattern shown in part (h) of Figure 1: as the baseline increases from zero to one, the conditional-on-baseline expected treatment effect must 1) start at zero, 2) decrease to negative values for lower baselines, 3) come back to zero, 4) go up to positive values for higher baselines, and then 5) descend back to zero as the baseline approaches one.

In addition to running reduced-form tests of these two predictions, we also perform a structural meta-analysis that allows us to estimate parameters of the data generating process of the experiments we analyze. This approach gives us additional results to assess our model. It additionally allows us to provide more precise advice for practitioners who are considering using information interventions in practice, as described in Section 6.

Throughout this section, we highlight the challenges that arise from attempting to test our model using data across existing experiments in the literature. However, we see our meta-analysis as the best way to test our model for two, related reasons. The first reason is that we wrote the model in part to rationalize existing experimental results from the literature. Testing the model on these data allows us to directly assess whether we have succeeded on this front. The second reason is that we aim to provide insights to practitioners who may want to deploy information interventions in various settings. If our model can successfully explain patterns of treatment effects across settings as diverse as the ones included in this meta-analysis, we can be more confident in its ability to guide practitioners in the diverse settings they will face.

This section proceeds as follows. Section 4.1 describes our approach to identifying papers and experiments to include in the meta-analysis. Section 4.2 provides details on the selected experiments and visually presents the data. Section 4.3.a presents reduced-form results assessing the first prediction, that negative treatment effects are more likely at lower baselines. Section 4.3.b reports on reduced-form results assessing our second prediction about the shape of the relationship between baseline and the magnitude and sign of the treatment effect. Section 4.4 introduces our structural

meta-analysis and presents estimates of key parameters in the literature.

4.1 Selecting Papers and Experiments for the Meta-Analysis

The first challenge in conducting a meta-analysis is identifying which studies to include. We developed a procedure to identify previously run experiments that were the most appropriate fits to our model (i.e., the setting was one modeled by the theory). Since our model made predictions about when we would see negative treatment effects and null results, we aimed to avoid selection based on "publication bias" or "file-drawer bias." We describe the procedure here.

First, we searched for papers that satisfied two criteria:

1. "At least one experimental treatment is attempting to influence a binary action by providing truthful information to subjects (e.g., telling subjects the % of others who take an action, telling subjects the benefits of taking a certain action, etc.)."[31,32]

2. "The paper reports the rate of taking the binary action in the control group, the treatment effect, and the standard error of the treatment effect (or these can be imputed)."

We found 18 papers (including working papers) that satisfied these criteria and then asked the experimental economics community to provide us with any additional papers that we had not identified.[33] We solicited papers with a request via email — with the subject line "Information Experiments (including file drawer)"—that we sent to the "ESA-discuss" mailing list.[34] The community sent us 25 papers in the period between when we sent that email and when we first presented our paper publicly, at which point

31. We only considered cases of experimenter-observed costly actions and excluded any experiments investigating hypothetical choices or self-reports, which ruled out papers such as Card et al. (2012), Kuziemko et al. (2015), and Karadja, Mollerstrom, and Seim (2017).

32. Note that continuous outcomes can be made into binary outcomes (e.g., "Is willingness to pay greater than $100?"), so the model can make predictions in such cases. However, determining the cutoff would provide a researcher degree of freedom that we wish to remove from the analysis.

33. This initial search was conducted by the authors of this manuscript and research assistants who were blind to the hypotheses for the meta-analysis that were generated by the model.

34. The email can be found at the end of the Appendix. The two criteria listed above are direct quotes from that email. ESA-discuss is the discussion e-mail list of the Economic Science Association. Following the rules of that mailing list, we included all of the papers we had found and invited individuals to send us any additional papers. Our email included 19 papers since we did not realize on a first read that Cialdini et al. (2006) did not provide rates of taking the binary action in the control group and thus did not qualify for inclusion in the meta-analysis.

we stopped accepting papers to include in the meta-analysis.[35] While we received many papers in response to our request, the vast majority were not appropriate fits of our theory (this may have been encouraged by the appeal in our email: "If you are unsure about whether to send a paper, please do so"), and we were only able to add 4 additional papers to our meta-analysis.[36]

In the Appendix, we list all the papers that we initially considered, or were sent to us in response to the solicitation and the main reason for exclusion (often papers were excluded for multiple reasons). The two most common reasons for a paper being excluded were that the treatment was not an information treatment or did more than just provide information (10 papers)[37] or the outcome reported by the authors was not binary (7 papers).[38] Papers were also excluded because they did not include a control group (1 paper) or because they were theoretical rather than empirical (1 paper). The final two papers were excluded because they violated the model for a more subtle reason: the information fully revealed the optimal action (in one case, the information was the value of an experimental asset, which fully revealed whether it should be bought or sold; in the other case, the information was that a mechanism was strategy proof, which fully revealed that truth telling was optimal). The decision to include or exclude a paper was made without looking at the paper's results.

Many of the papers included in the meta-analysis include multiple experiments that qualify based on our inclusion criteria (e.g., if multiple information interventions are being tested in the same setting). Consequently, we are able to analyze the results of 75 experiments from the 22 papers we identify. Each experiment includes a baseline and a treatment effect, and these are the data that we use to perform our statistical tests. Note that when experimental treatments are run as separate arms in the same intervention, they will have the same baseline in our data since they share

35. We first presented the paper publicly on August 22, 2016 at the Experimental Economics session of the Stanford Institute of Theoretical Economics. At that point, we stopped accepting papers to avoid any potential selection of papers into our meta-analysis. For example, this could introduce positive selection whereby individuals who knew the results of the model might send us papers that were consistent with its predictions.

36. We take this as a sign that we had successfully identified the majority of relevant papers in our first pass of collecting relevant work.

37. For examples: if treatments provided information but also changed the strategic structure of the game, then we excluded that paper; if multiple interventions were run, and the effect of information could only be estimated by assuming no interaction with other treatments, then we excluded that paper; if the treatment was advice from another subject rather than a nudge provided by the experimenter, we excluded the paper.

38. We included cases where the authors reported both a binary outcome (e.g., whether a student graduated high school, whether a subject donated anything in a dictator game) alongside a continuous outcome (e.g., the amount of schooling achieved, and the amount donated). In cases where multiple binary outcomes were reported from the same information intervention, we used whichever outcome was the primary focus of the author(s). Note that we exclude papers that only include continuous variables, such as Nguyen (2008).

a control group.

Before we present the results of the meta-analysis, we highlight three potential concerns with using papers collected from the literature to test our theory. One concern is that we are aiming to explain experiments that differ across many dimensions (i.e., they vary in terms of settings and in terms of the information interventions provided).[39] As described above, we see this as a "feature" rather than a "bug" of our meta-analysis. In addition, the variation works against us finding anything systematic and so makes us more confident if we are able to succeed in rationalizing these results.

A second concern is that we might not identify all relevant experiments in the literature due to a "publication bias" (where certain experiments never get published) or a "file-drawer bias" (where certain experiments are never written up into shareable manuscripts). That is, despite specifically asking for research "including file drawer" in our solicitation, it is possible that certain experiments (e.g., experiments with insignificant or negative treatment effects) were not readily available, perhaps never written up as a manuscript. Missing these experiments, if they indeed exist, provides us with less data, which would decrease our statistical power. A more troubling concern, however, would arise if "missing" experiments displayed a systematic relationship between treatment effect and baseline that might bias our estimate of their relationship. However, we are unable to construct any reasonable explanation for why publication bias or file-drawer bias would lead disproportionately negative treatment effects to appear missing from certain ranges of baselines (but not others).[40] Finally, in our data, a substantial portion of the effect sizes are negative and/or very small in magnitude: For example, 24.0% of the treatment effects are negative, and 50.7% have an absolute value less than or equal to 0.25.

Finally, as with all meta-analyses, any across-paper variation in our sample is not random. Specifically, there may be other features that covary with baseline rates across papers. While finding support for the predictions made by our model, especially the specific down–up–down pattern, will be convincing, it is impossible for such an endeavor to be dispositive.

39. Our 75 experiments include lab experiments, field experiments, and framed field studies. Outcomes vary dramatically, with examples including: paying taxes, reusing a hotel towel, continuing in secondary schooling, ordering a popular dish at a restaurant, and deciding to join Teach For America as a teacher.

40. The only plausible concern we can envision in this regard is due to binary variables having higher variance at intermediate values. If researchers do not properly respond to an expected increase in variance with a proper increase in sample size, and choose not to write-up papers with null results, we might expect to be missing experiments with intermediate baselines (e.g., close to 0.5) with relatively small (positive or negative) treatment effects. This would create a hump-shaped curve. It would not create an area with negative treatment effects. We will keep this potential concern in mind when analyzing the results, below.

4.2 Descriptive Statistics

Figure 2 presents an unstructured view of the data. Each experiment is represented by a dot on the graph, where mean baseline take-up is shown on the x-axis and the reported average treatment effect is shown on the y-axis.[41] In all the experiments we analyze, the information intervention was designed to increase take-up of an action, so positive treatment effects indicate a treatment effect in the intended direction, and all negative treatment effects indicate a backfire.

Before testing predictions of the model, we document a few observations from the data. First, across all experiments in our meta-analysis, the average treatment effect is 0.02, with a median of 0.01; the typical information intervention has a modest positive effect on behavior. Second, most experiments are done on low baseline rates: the median baseline take-up across all experiments is 0.34, and 73.0% of all experiments have a baseline of 0.25or less.[42] These statistics are consistent with the folk intuition that information interventions should be tried when baseline is low (so that a large share of agents are available to be induced to take the desired action). We provide further evidence of this folk intuition in Section 5.

4.2.a Naïve Analysis of Information Interventions

Before we test predictions of the model, we consider what the data would suggest absent the insights in the theory. In short, if a researcher or policymaker, who was naive to our model, looked at our data, what might they conclude about the effectiveness of information interventions?

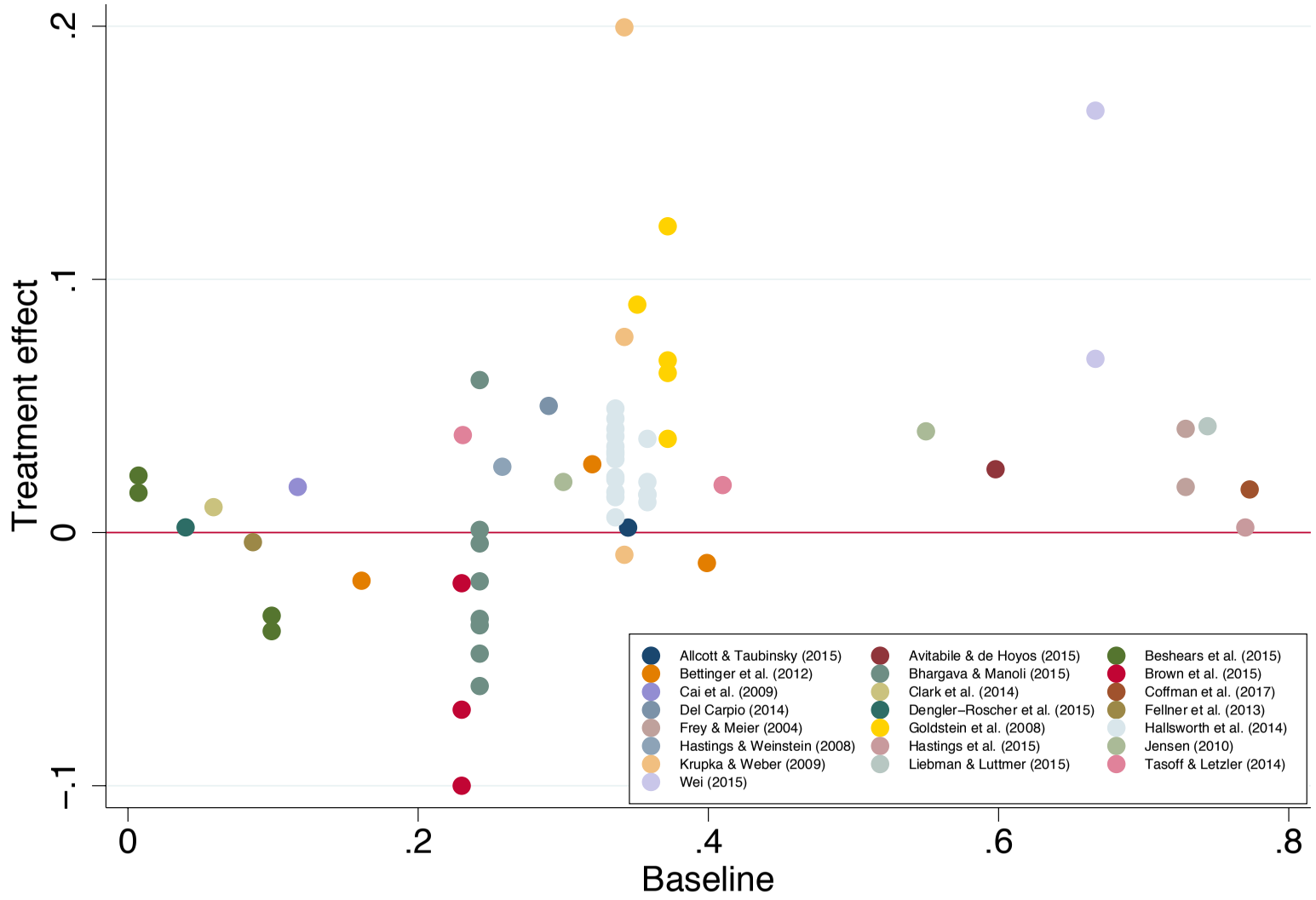
Not taking into account the importance of baseline rate, across the experiments in our metadata, the results are modest. There is almost a quarter of a chance of the intervention backfiring: 24.0% of treatment effects are negative. Most effects are small. 58.7%, have a treatment effect less than or equal to 0.02. As mentioned above, the average treatment effect is 0.02. Further, few treatment effects are positive and large. Only 14.7% of treatment effects are greater than or equal to 0.05.

A researcher or practitioner looking at such numbers might be discouraged. As a result, they might not utilize an information intervention for their upcoming project. However, this analysis masks important heterogeneity.

In light of our model, we see that information can be a very effective policy tool. Indeed, in these experiments, information is working, and working

41. To avoid visually compressing the bulk of the data, 3 experiments with treatment effects greater than 20 percentage points are not included in the figure. This exclusion is for the usefulness of the visual representation of the data only, and all experiments are included in the analysis that follows.

42. Note that baseline rates are typically from control groups where contact has been made (e.g., a letter was sent, but the experimental information was not provided). In that way, the treatment effects are the effect of providing the information as opposed to, for example, the effect of sending a letter and providing information.



Three observations from Krupka & Weber (2009) and one from Brown et al. (2015) have treatment effects outside the range of the vertical axis.

Figure 2: Meta-Analysis Data

in the way predicted by our model.

4.3 Reduced-Form Analysis

Our model makes two empirical predictions: that the probability of positive treatment effects increases in baseline take-up rate, and that the expected treatment effect as a function of baseline follows a "down–up–down" relationship. We will first consider these predictions in turn.

4.3.a Positive Treatment Effects Are More Likely at Higher Baselines

The relation between the probability of positive treatment effects and the baseline is suggested visually in Figure 2. The negative treatment effects are generally found at low baselines while the positive treatment effects are more common at high baselines. For example, 45.8% of treatment effects are negative when the baseline rate is below 0.25; however, only 12.2% are negative for baselines above 0.25. Table 1 regresses whether a treatment effect is positive on baseline, using a linear probability model (Column 1) and a probit (Column 2). Standard errors are provided by a two-tier hierarchical bootstrap.[43] First, to weight by the precision of each experiment's results, the bootstrap resamples at the experimental treatment cell: In each iteration of the bootstrap, we create new data for every individual treatment and control cell by drawing, with replacement, from the original data in that cell until the original sample size is reached. Second, to account for potential within-paper correlations, the bootstrap then resamples which papers are included, again drawing with replacement, using the redrawn data from the first step for each experiment within a paper.

Both specifications in Table 1 estimate a significant positive relationship between baseline and the probability of a positive treatment effect (0.08 of bootstrapped OLS estimates and 0.04 of bootstrapped probit estimates are non-positive).[44] The magnitudes of the estimate are quite large. In Column 1, every 0.1 increase in the baseline, implies a 6.8 percentage point increase in the likelihood of a positive treatment effect. It is also worth noting that the adjusted-R^2 of the linear probability model is 0.08, with a pseudo-R^2 of 0.10 for the probit. That is, knowing *only the baseline rate*, the linear model can predict 8–10% of the variation in the sign of the treatment effect across papers in our sample.[45] Noting that these papers vary in many

43. Most of our bootstrapped coefficient distributions are non-normal (with a thick right tail), so for standard errors, we report half the width of the interval centered around the median that contains 68.27% of the data. (For a normal distribution, 68.27% of the data lies within one standard deviation of the median.)

44. The p-values implied by the frequency approach in the bootstrap differ from calculating z-statistics from the point estimates and standard errors because the bootstrapped distributions are both asymmetric, with thick right tails. Since the distributions are non-normal, we do not rely on z-statistics for p-values.

45. As noted above, these papers vary dramatically with respect to context, population, outcome, and information provided in the treatment. We find it striking that the model has

Table 1: Predicting Likelihood of Positive Treatment Effects with Baseline Rates

	OLS	Probit
	(I)	(II)
Baseline	0.68	0.86
s.e.	(0.34)	(0.47)
p-value	[0.08]	[0.04]
Constant	0.76	0.76
s.e.	(0.10)	(0.10)
p-value	[0.00]	[0.00]
Papers	22	22
Studies	75	75
Adjusted/pseudo-R^2	0.08	0.10

NOTES. Table shows estimates regressing whether treatment effect is positive on baseline rate. Column (I) presents linear probability model and Column (II) Probit. Standard errors are the standard deviation of estimates from hierarchical bootstrap described in Section 4.3.a. Note both boot-strapped distributions of estimates are right-skewed. "p-value" reports the percentage of estimates from the hierarchical bootstrap described in Section 4.3.a that are non-positive. De-meaned baseline rates used, so constant reports estimate at the average baseline rate.

important ways—lab or field, large stakes or small stakes, social information or direct information—it is impressive that only the baseline rate can predict so much of the variance in the outcomes.

4.3.b Treatment Effect Versus Baseline Has a "Down–Up–Down" Shape

The second prediction of our model that we test is the "down–up–down" relationship between baseline rates and treatment effects: As baseline increases from 0 to 1, treatment effects decrease from 0, increase through 0 (producing a small negative "nub"), increase past 0, and decrease back to 0 (producing a large positive "bump"). This theoretical relationship is shown in Figure 1h.

One challenge to testing this prediction is that, while we expect any given experiment to have a theoretical down–up–down curve, different experiments are expected to have different curves. The baselines at which the treatment effect curve hits its minimum and maximum, and the baseline where the curve has its intermediate 0, will vary depending on the context (e.g., the strength of agents' priors, where the nudge falls relative

such predictive power while blind to all parameters other than baseline take-up.

to those priors, and the distribution of priors relative to the distribution of thresholds). Given this underlying variation, testing for this down–up–down pattern is a challenge. We see this exercise as a stress test of how well our model fits the data.

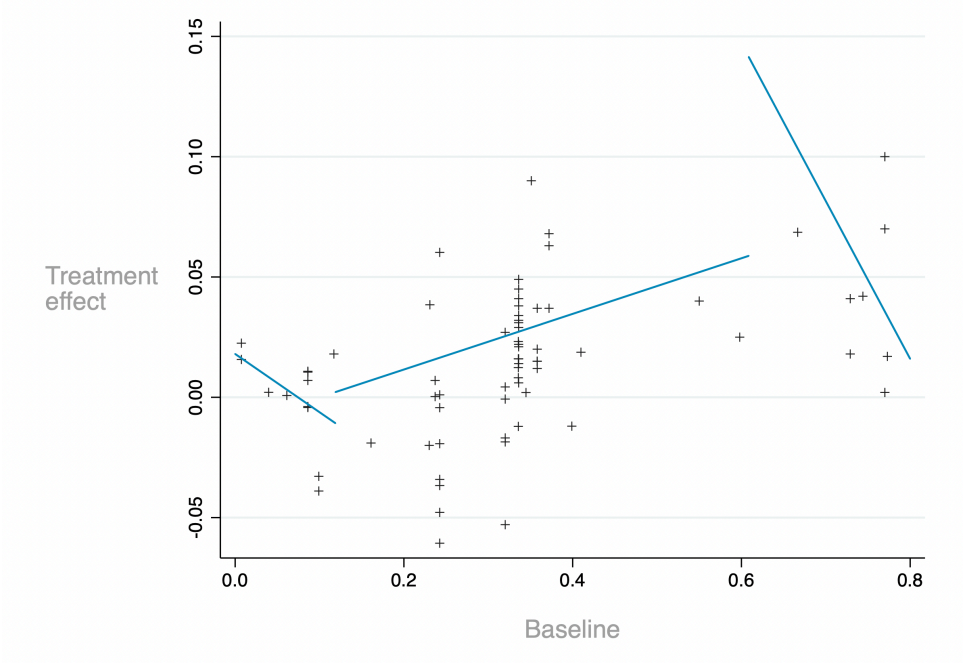
Similar to Simonsohn (2018), which proposes testing a hump-shaped or U-shaped relationship using a piece-wise linear test of two lines, we test for a down–up–down relationship using three lines: we test if the slope of the leftmost line is negative, the second positive, and the third negative, i.e., down–up–down.[46] The challenge of this method is to choose the break points between the three lines in a way that the analysis is well situated to detect the proposed pattern if it exists, but it does not introduce a researcher degree of freedom.

To these ends, it seems most natural to begin with break points at the estimated local minimum and local maximum and to consider other break points nearby as well. That is, from an OLS-fitted third-degree polynomial (unreported), we will estimate the baselines where the local minimum and local maximum occur, and use these as the thresholds separating our three lines. These local extrema occur roughly at baseline rates of 0.119 and 0.609. We consider these, and nearby, break points.

Figure 3 shows a scatter plot of studies in our sample with the three fitted lines overlaid. First, note visually the break points appear reasonable. If one were trying to find a down–up–down pattern, they would likely select break points in the vicinity of those chosen by the cubic fit. Second, all three lines are directionally consistent with predictions (Also see Table 2): the first is downward-sloping with an estimated slope of -0.24 ($p = 0.045$).[47] The second is upward with a 0.12 slope ($p = 0.14$). The third is downward-sloping with an estimated 0.66 slope ($p = 0.19$); however, only the first line is significant by conventional standards. We can get a better sense of the veracity of the pattern in two ways. First, we can redo the analysis removing any data that have outsize influence on the estimates, i.e. outliers. To find such leveraged data, we estimate how much the coefficients change when each data point is individually removed. We find one data point decreases the estimates of the second line by almost a full standard error range (-0.98).

46. Following Simonsohn (2018), we allow for discontinuities at those break points. That is, we estimate a separate regression for each of the three regions.

47. The p-values for all three lines in this section, and in Table 2 and Appendix Table A.1, come from a bootstrap that re-samples both papers and experimental data. We utilize the same hierarchical bootstrap used in Section 4.3.a making one accommodation for the three-line analysis. When one of the three lines only has a handful of papers in its data set, re-drawing papers often leaves only one or two papers for the analysis of that line. To avoid these unhelpful estimations, we instead utilize a jackknife procedure for resampling papers: We remove one paper, redo the analysis, re-include that paper, remove the next, and so on. When each paper is removed, the data within each experimental cell is redrawn i.i.d. with replacement, as we did in Section 4.3.a. We redraw the data 1,000 times for each combination of break-points and paper removal. Each p-value is the percentage of estimates inconsistent with the model's prediction.



Crosses each represent a study and its baseline rate - treatment effect pair. Blue lines show fitted OLS estimates for baseline rates below 0.119, between 0.119 and 0.609, and above 0.609 respectively. 0.119 and 0.609 are the baseline rates where the fitted cubic is estimated to be at its local minimum and maximum respectively. For visual simplicity treatment effects over 0.1 excluded in the scatter plot, but they are included in the estimation of the lines.

Figure 3: Three-line fit predicting treatment effect with baseline rate

Table 2: Estimating slopes for three lines

	First Line	Second Line	Second Line	Third Line
		All Data	Excl. Outlier†	
	(I)	(II)	(III)	(IV)
Baseline	−0.24	0.12	0.21	−0.66
s.e	(0.10)	(0.10)	(0.14)	(0.40)
p-value	[0.05]	[0.14]	[0.02]	[0.19]
Papers	5	12	12	6
Studies	12	54	53	9
Adjusted-R^2	0.13	−0.00	0.06	0.20

NOTES. Table shows estimates regressing treatment effect on baseline rate. First Line presents estimates for baseline rates less than 0.119. Second Line presents estimates for baseline rates greater than 0.119 and less than 0.609. Third Line presents estimates for baseline rates greater than 0.609. Standard errors are the standard deviation of estimates from hierarchical bootstrap described in Section 4.3.a modified as described in Section 4.3.b, footnote 47. Note the boot-strapped distributions of estimates are skewed, left-skewed for (I), (III), (IV), and (VI), and right-skewed for (II) and (V). "p-value" reports the percentage of bootstrapped estimates that are inconsistent with the model's prediction: i.e. non-negative for (I), (III), (IV), and (VI) and non-positive for (II) and (V). †One extreme outlier removed as described in Section 4.3.b (Removal only affects the second line's estimate).

The second and third largest influence measures, in absolute terms, are 0.36 and 0.21. Though admittedly ex post, we redo the analysis removing the most extreme outlier.[48] Column (III) of Table 2 shows this new analysis. The second line is upward-sloping with estimated 0.21 slope, with $p = 0.02$. That is, once we remove a data point that has by far the largest effect on our analysis, and look for a pattern in the rest of the data, the analysis is more consistent with the model's predictions.

Second, to further test the robustness of the findings, we redo the analysis using other break points, just above and below those already used. We move the break points to the left and right, including or excluding at least two more data points. We only use break points that include at least four data points in the first and third lines. Altogether, we test four break points between the first and second line and four break points between the second and third lines. This provides four estimates for the first line, sixteen for

48. The data point removed comes from a laboratory experiment with a baseline rate of 0.23 and a treatment effect of 0.29 (Brown, Trautmann, and Vlahu 2017).

the second line, and four for the third line. All sixteen regressions, and all break points, can be found in Appendix Table A.1. For the first line, three of the four estimates are negative. The lone positive estimate is the only instance in which the extreme outlier discussed above is included in the first line. The p-values for the three negative lines are 0.00, 0.045, and 0.045. All sixteen estimates for the second line are positive, i.e. consistent with the model, with an average estimate of 0.18, and an average p-value of 0.06. Our meta dataset does not include a wealth of studies with high baseline rates, and the third line is noisy as a result. Two of the four estimates are negative, i.e. consistent with the model, though they are the two with the most data. When only four or six data are used to estimate the third line, estimates are positive and noisy.

Taken together, our reduced form results suggest that the experiments in the data may follow the specific down–up–down pattern predicted by the model. We find some evidence for this pattern despite significant variation in settings, outcomes, and information interventions across the experiments in our meta-analysis, providing additional credibility of the explanatory power of our model.

4.4 Structural Meta-Analysis

The results above offer a reduced-form approach to testing two hypotheses from the model: negative treatment effects will disproportionately appear at low baselines and the treatment effect will follow a down–up–down pattern with respect to baseline. However, the reduced-form approach leaves open a few important questions. In this section, we will introduce a specific instance of the more general model introduced in Section 3.3. We will then structurally estimate key parameters of that model.

That so many experiments were attempted at low baselines allowed us to claim that practitioners follow the intuition to use information nudges in settings with a low baseline (since many people are available to be nudged). The structural approach will allow us to assess whether practitioners also follow the intuition to provide nudges that are good news on average by providing an estimate of where these practitioner's nudges fell in agents' prior belief distributions.

In the reduced-form analysis above, we posited significant underlying variation across experiments, but we had no way of quantifying this variation. The structural approach will identify the variation in two key parameters that differ across experiments: where a nudge falls in the agents' prior belief distribution and the relative variation in agents' thresholds and beliefs in a given experimental setting. Once we perform the structural analysis, we can investigate how much variation in treatment effect curves we should expect across experimental settings in practice.

Finally, the structural approach provides guidance to practitioners as they consider potential nudges by allowing them to assess their likely ef-

fects by comparing their potential nudge to nudges used previously in the literature.

4.4.a No literature-level noise

The literature-level model from Section 3.3 assumes that priors are drawn from the standard normal and that thresholds are drawn independently from some family of distributions indexed by β, the baseline. We choose normal distributions with mean $-\Phi^{-1}(\beta)\sqrt{1+\eta^2}$ and variance η^2, i.e., $\Theta \sim \mathcal{N}\left(-\Phi^{-1}(\beta)\sqrt{1+\eta^2}, \eta^2\right)$. Intuitively, this means that thresholds are normally distributed with a standard deviation η times larger than that of priors and with a mean that makes the baseline equal to β.[49] For now, we assume the threshold width, η, remains constant across experiments. We will also assume that within an experiment, the update strength is independent of thresholds and priors. As with the threshold width, across experiments, we assume that the expected update strength, $\mathbb{E}[E]$, remains constant. (In subsequent sections, we will allow for cross-experiment heterogeneity in η and $\mathbb{E}[E]$.) We complete our model by assuming that the nudge signal is distributed independently and normally across experiments, with mean $\bar{\nu}$ and variance σ_ν^2, i.e., $N \sim \mathcal{N}(\bar{\nu}, \sigma_\nu^2)$.

Broadly speaking, we identify an experiment in the literature as a realization, (ν, β), of the random vector that represents the nudge signal and baseline, (N, B). In other words, using the terminology introduced in Section 3.3.c, there is no *literature-level noise*.

Given this model, the treatment effect is positive when the nudge signal, ν, exceeds the expected prior among marginal agents, $\mathbb{E}_\beta[M \,|\, \Delta = 0]$.[50] From here, we can do a bit of algebra and derive the conditional-on-baseline distribution of the treatment effect.

> **Result.** *For the parametric model of this section, when the realization of (N, B) is (ν, β), the treatment effect is given by*
>
> $$\tau_{\beta,\nu} = \varphi\left(\Phi^{-1}(\beta)\right)\left\{\frac{\mathbb{E}[E]}{\sqrt{1+\eta^2}}\,\nu + \frac{\mathbb{E}[E]}{1+\eta^2}\,\Phi^{-1}(\beta)\right\}.$$
>
> *Since N is distributed normally with mean $\bar{\nu}$ and variance σ_ν^2, this means that, conditional on the baseline, B, having realization β, the treatment effect is normally distributed, with mean and standard devi-*

49. To see this, note that since $\Delta = \Theta - M$, the rules of normal distributions dictate that $\Delta \sim \mathcal{N}(\tilde{\theta}, 1+\eta^2)$, where $\tilde{\theta} \equiv -\Phi^{-1}(\beta)\sqrt{1+\eta^2}$. Hence the fraction of agents that take up (i.e., the fraction with $\Delta \leq 0$) is $\Phi\left(-\tilde{\theta}/\sqrt{1+\eta^2}\right)$, which is equal to β.

50. Here, Expression 7 simplifies because update strengths are independent of thresholds and priors.

ation given by

$$\mathbb{E}\left[\tau_{B,N} \,\middle|\, B=\beta\right] = \varphi\left(\Phi^{-1}(\beta)\right)\left\{\frac{\mathbb{E}\left[E\right]}{\sqrt{1+\eta^2}}\,\bar{\nu} \;+\; \frac{\mathbb{E}\left[E\right]}{1+\eta^2}\,\Phi^{-1}(\beta)\right\},$$

$$\sqrt{\mathrm{Var}\left[\tau_{B,N} \,\middle|\, B=\beta\right]} = \varphi\left(\Phi^{-1}(\beta)\right)\frac{\mathbb{E}\left[E\right]\sigma_{\nu}}{\sqrt{1+\eta^2}}. \tag{11}$$

From this, it almost immediately follows that the conditional-on-baseline probability of a positive treatment effect is given by

$$\Pr\left\{\,\tau_{B,N}>0 \;\middle|\; B=\beta\,\right\} \;=\; \Phi\left(\frac{\mathbb{E}\left[\tau_{B,N} \,\middle|\, B=\beta\right]}{\sqrt{\mathrm{Var}\left[\tau_{B,N} \,\middle|\, B=\beta\right]}}\right). \tag{12}$$

The proofs for these results can be found in Section A.3.a of the appendix.

An immediate corollary shows how this model can be estimated: we divide the treatment effect by $\varphi\left(\Phi^{-1}(\beta)\right)$ and regress on $\Phi^{-1}(\beta)$. That is,

Corollary. *If we regress $\tau_{\beta,N}/\varphi\left(\Phi^{-1}(\beta)\right)$ on $\Phi^{-1}(\beta)$, assuming homoskedastic errors, we are estimating the equation*

$$\frac{\tau_{\beta,N}}{\varphi\left(\Phi^{-1}(\beta)\right)} \;=\; a \;+\; b\,\Phi^{-1}(\beta) \;+\; \zeta, \tag{13}$$

where a and b are constants defined by

$$a \equiv \frac{\mathbb{E}\left[E\right]\bar{\nu}}{\sqrt{1+\eta^2}} \qquad \textit{and} \qquad b \equiv \frac{\mathbb{E}\left[E\right]}{1+\eta^2},$$

and ζ is a normally distributed error with mean zero and standard deviation defined by

$$\sigma_{\zeta} \equiv \frac{\mathbb{E}\left[E\right]\sigma_{\nu}}{\sqrt{1+\eta^2}}.$$

The estimates for this regression in the full sample are listed in the "Full sample" column of Table 3. Unfortunately, the estimate for parameter b is negative, which is not theoretically possible in this section's model. In Figure 4, we show a scatter plot for the regression. The full sample includes both the darker crosses and the lighter, labeled, circular points; the lighter, negatively sloped, dashed line is the full-sample linear fit.

Clearly, the lighter points (whose source papers are labeled in Figure 4) have much higher values of $\tau/\varphi\left(\Phi^{-1}(\beta)\right)$ than the rest of the sample. In fact, the mean dependent-variable value of those four studies is about 19 times larger than that of the the other studies. Essentially, we seem to

Table 3: Regression estimates for Equation 13

	OLS		Robust regression
	Full sample	Outliers excluded	Stata `rreg` command
$a \equiv \frac{\mathbb{E}[E]\ \bar{\nu}}{\sqrt{1+\eta^2}}$	0.0716 (0.0338)	0.0935 (0.0155)	0.0822 (0.0150)
$b \equiv \frac{\mathbb{E}[E]}{1+\eta^2}$	−0.0468 (0.0418)	0.0998 (0.0217)	0.0823 (0.0197)
$\sigma_\zeta \equiv \frac{\mathbb{E}[E]\ \sigma_\nu}{\sqrt{1+\eta^2}}$	0.2274 (0.0186)	0.1013 (0.0085)	0.1002 (0.0083)
Papers	22	21	21
Studies	75	71	71
R^2	0.02	0.23	0.19
Goodness-of-fit p-value	0.273	0.00002	0.00008

NOTES. The left and center columns correspond to the dashed and solid lines in the figure below. The rightmost column reports the results of a more sophisticated (but harder to interpret) approach to outliers that follows Li (1985), as described in footnote 51. Goodness-of-fit p-values are assessed using F-tests. Standard errors are in parentheses.

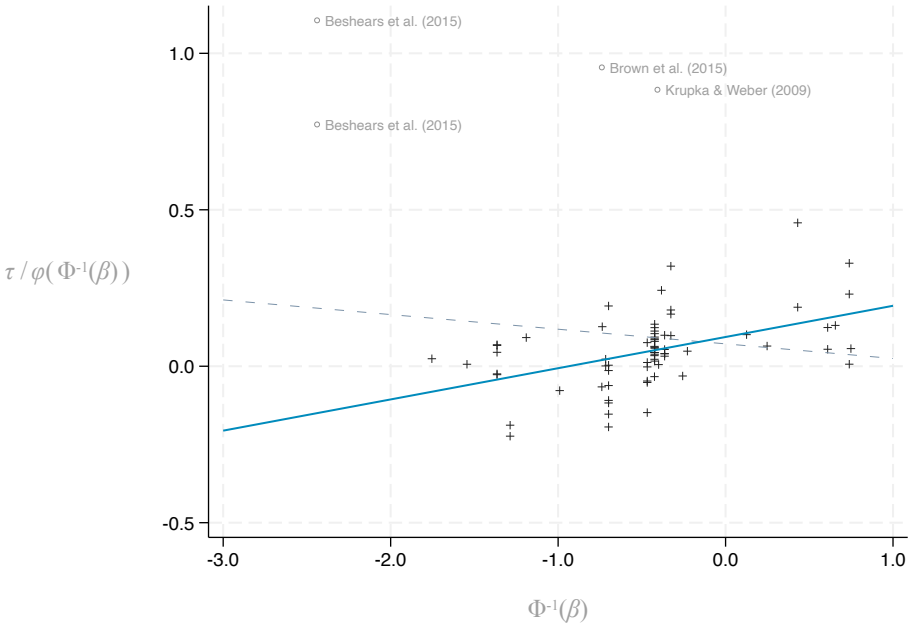


Figure 4: Linear fits for Equation 13

NOTES. Each study in the sample is given a darker cross or a lighter circle. The circles are studies we throw out as outliers: their dependent-variable—i.e., $\tau/\varphi(\Phi^{-1}(\beta))$—is, on average, 19 times larger than that of the other studies. The outliers have their sources studies written next to them. The light, dashed line is the best linear fit across the entire sample; the dark, solid line is the best linear fit with outliers excluded.

have an outlier problem. In Section A.3.b of the appendix, we make a more technical case that these four studies are indeed outliers, but in the main text, we limit ourselves to a simple visual inspection of the data.

Estimates for our regression that exclude the four outliers are listed in the "Outliers excluded" column of Table 3. Now, the parameter estimates are all positive and the regression itself is highly significant. Its R^2 is 0.23, which means that across the wide variety of contexts present in our dataset, the simple regression described by Equation 13 explains 23% of the variance in the data, once outliers have been removed.

To provide some additional support for our rather simple outlier-mitigation protocol, in Table 3, we also report "Robust regression" estimates using the more sophisticated (but harder to interpret) methods used by Stata's `rreg` command.[51] The `rreg` command throws out the same four outliers and gives results that aren't significantly different from the simpler "Outliers excluded" results. As such, we will discuss the "Outliers excluded" results in what follows.

Begin by noting that the ratio a/σ_ζ, whose value is 0.92 (with a standard error of 0.17), is an estimate of $\bar{\nu}/\sigma_\nu$. That the ratio is significantly positive tells us that the mean of the nudge-signal distribution is higher than the mean of the prior distribution. This makes sense: experimenters are trying to run information interventions where the signal is good news to the target population as a whole. In fact, we can go a bit further: in our model, $\Phi(\bar{\nu}/\sigma_\nu)$ is the probability that the nudge signal exceeds the mean of the prior distribution. This probability computes out to 82.2% (with a standard error of 4.5 percentage points). Again, this lines up well with our intuition concerning the sorts of information interventions experimenters like to run: more than 4/5 of studies have a nudge signal that exceeds the mean of the prior distribution.

We can also look at the ratio a/b, which is an estimate of $\bar{\nu}\sqrt{1+\eta^2}$. Its estimated value is 0.93 (with a standard error of 0.16). Since $\sqrt{1+\eta^2}$ is bounded below by one, we can then infer that $\bar{\nu}$ is bounded above by 0.93. Interpreted against the distribution of priors, this tells us that the mean signal is between the 50th percentile of the prior distribution (since its estimate is positive) and the 83rd percentile of the prior distribution (since 0.93 is at the 83rd percentile of the standard normal). Again, this lines up well with our intuition concerning the sorts of information interventions experimenters like to run.

In addition to this sort of analysis, we can also use the formulas in Equations A.15 and A.16 to compute structural estimates of the expected treatment effect and the probability of a positive treatment effect and as a func-

51. Essentially, `rreg` follows a procedure described in Li (1985) that involves screening by Cook's D (Cook 1977) and then computing a Huber M-estimator (Huber and Ronchetti 2011) via iteratively reweighted least squares, using Huber weights (Huber 1964) to find a starting point and then Tukey biweights (Beaton and Tukey 1974) to converge from there.

tion of the baseline. These predictions are plotted out in Figure 5.

Figure 5a plots out the probability of a positive treatment effect as a function of baseline. More specifically, it is a plot of Equation A.16 with the estimates from Table 3 plugged in. Looking at the plot, we see that there is a roughly 75% chance of a backfire when the baseline is 5%. This drops to a 50% chance of a backfire when the baseline increases to 17% and a 25% chance of a backfire when the baseline increases to 40%. It isn't until the baseline hits 64% that the chance of a backfire gets pushed below 10%. Across all experiments in our outliers-excluded sample, the mean predicted probability of a backfire is 34%. So, while backfires aren't the most likely outcome, they are far from uncommon.

Now, we move on to the predicted treatment effect as a function of the baseline, which is plotted out as the solid line in Figure 5b. It is computed by plugging the estimates from Table 3 into Equation A.15. The qualitative shape of the curve is similar to that seen in part (h) of Figure 1. The dashed lines in Figure 5b are plus/minus one standard deviation. (Again, this is computed by plugging the estimates from Table 3 into Equation A.15.) That is, *the dashed lines do not indicate a standard error of measurement*. Rather, they indicate the spread around the expectation that is predicted by our model. So, while we predict the expected treatment effect to vary in a certain way, we should also expect a good deal of noise around the trend. This is unsurprising given the breadth of applications the model is designed to capture.

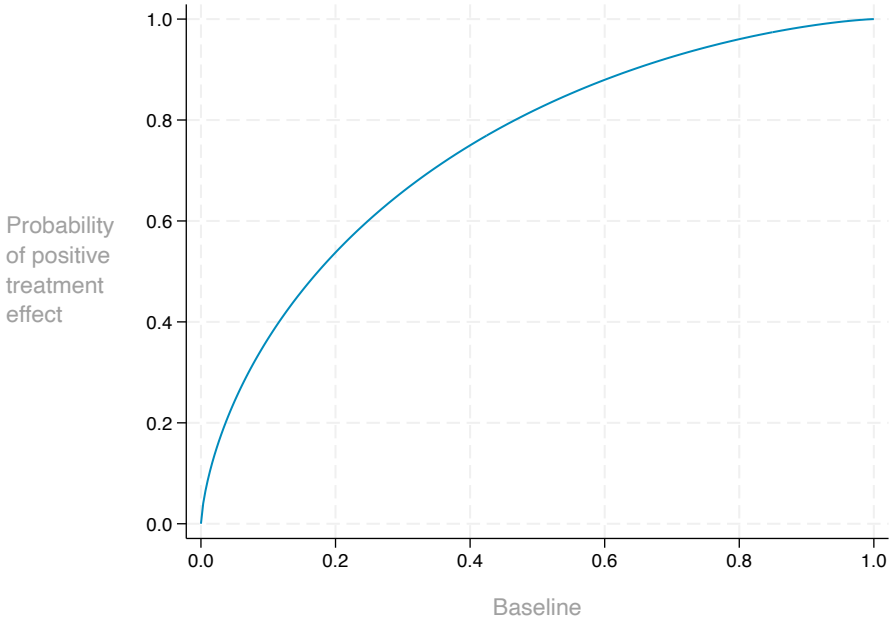
The expected treatment effect is negative for baselines below about 17%; however, backfires are within one standard deviation of the expected outcome for baselines all the way up to about 53%. This corroborates the narrative we discussed concerning the predicted probability of a positive treatment effect: while backfires might not be the most likely outcomes, they are far from uncommon.

In addition to showing the prevalence of backfires, Figure 5b also shows that there are "sweet spots" for big negative and big positive treatment effects. When baselines are in the 60–85% range, the treatment effects the most positive, and when baselines are in the 3–9% range, treatment effects are the most negative.[52]

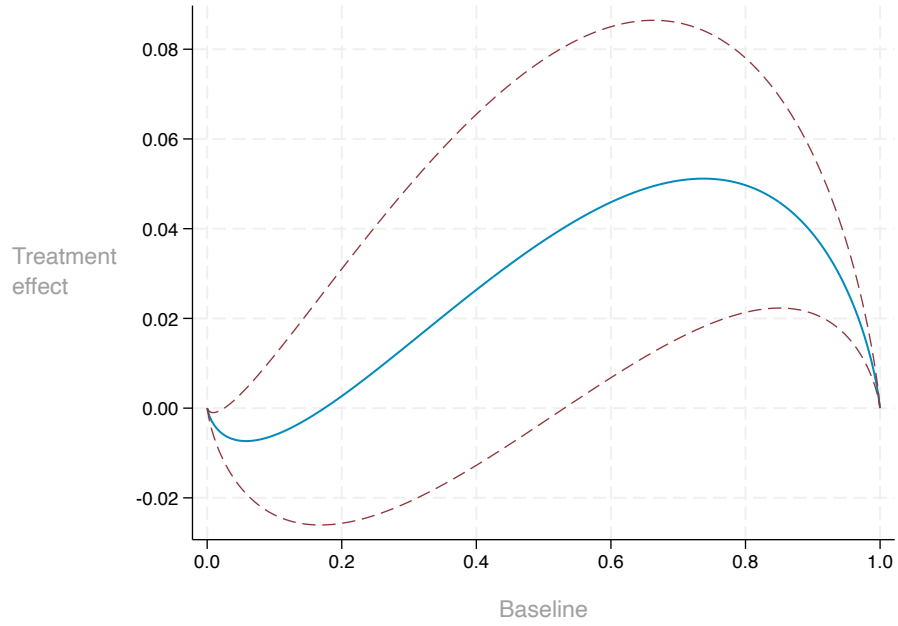
While the plots in Figure 5 are quite useful, we have not yet been able to estimate the individual parameters in our model, as they are not identified by the likelihood defined in Equation 13.[53] In the next section, we solve this identification problem by modeling cross-experiment heterogeneity in η and $\mathbb{E}[E]$. In other words, using the terminology introduced in Sec-

52. These ranges were computed by looking for the baseline ranges where the expected treatment effect is greater than 90% of its maximum value or less than 90% of its minimum value.

53. To see this, note the same values for a, b, and σ_ζ are attained when either $\left(\bar{\nu}, \sigma_\nu, \eta, \mathbb{E}[E]\right)$ or $\left(\bar{\nu}\sqrt{1+\eta^2},\ \sigma_\nu\sqrt{1+\eta^2},\ 0,\ \mathbb{E}[E]/\left(1+\eta^2\right)\right)$ is plugged into their definitions.



(a) The predicted probability of a positive treatment effect



(b) The predicted treatment-effect expectation and standard deviation

Figure 5: Predictions from the Equation 13 regression

NOTES: In subfigure (a), the predicted probability of a positive treatment effect is plotted by plugging the estimates from Table 3 into Equation A.16. In subfigure (b), the solid line is a plot of the expected treatment effect, while the dashed lines are plus/minus one standard deviation. All lines comes from plugging the estimates from Table 3 into Equation A.15. To be clear, the dashed lines are not measurement error, but rather the spread in outcomes predicted by our model fit.

tion 3.3.c, we will introduce some literature-level noise.

4.4.b Introducing literature-level noise

The model from Section 3.3 doesn't incorporate literature-level noise: a random experiment in the literature is determined entirely by the baseline and the nudge signal drawn. This is how the model from the previous subsection works, since η^2 and $\mathbb{E}[E]$ are treated as constant across all experiments. In this subsection, we will introduce literature-level noise by allowing cross-experiment heterogeneity in η. Doing so will allow us to better identify the individual parameters in our model.

Since η^2 is a variance, it should have support on the positive real line; hence, we model it as the realization of an independently drawn, gamma-distributed random variable, H^2. Since $\mathbb{E}[E]$ is a convex weight, it should have support on the unit interval; hence, we model it as the realization of an independently drawn, beta-distributed random variable, $\mathcal{E}$. To summarize then, for a given baseline, the experiment is now dependent on the draw of $(N,\ H^2,\ \mathcal{E})$, where all three random variables are independent and distributed according to

$$\begin{aligned} N &\sim \mathcal{N}(\bar{v},\ \sigma_v^2), \\ H^2 &\sim \text{Gamma}(k,\ t), \\ \mathcal{E} &\sim \text{Beta}(b,\ g). \end{aligned}$$

Note we are using the shape–scale parametrization of the gamma distribution, whose mean and variance are $k\,t$ and $k\,t^2$, and the modified PERT parametrization of the beta distribution (Clark 1962), whose mean and *mode* are $(1+b\,g)/(2+g)$ and b.[54] (In the large g limit, the expectation converges to the mode, which intuitively suggests that $1/g$ is a proxy for variance.) Further note that k, t, and g must be positive, while b must lie on the unit interval.

Then, the likelihood of observing treatment effect τ when the baseline

54. The more common beta parametrization is in terms of the two parameters, α and β, where the mean is $\alpha/(\alpha+\beta)$. In terms of this parametrization, the modified PERT parametrization can be expressed as $\alpha = 1 + b\,g$ and $\beta = 1 + (1-b)\,g$ Note that the modified PERT parametrization has the added benefit of ruling out beta distributions with both α and β less than one. Such distributions are U-shaped (i.e., bimodal at zero and one), which would be an odd choice for our particular application.

is β is given by

$$\Pr\{\tau \mid B = \beta\}$$
$$= \int_0^1 \left\{ \int_0^\infty \frac{1}{\mathbb{E}[E]\,\sigma_\nu/\sqrt{1+\eta^2}}\, \varphi\left(\frac{\frac{\tau}{\varphi\left(\Phi^{-1}(\beta)\right)} - \frac{\mathbb{E}[E]\,\bar{\nu}}{\sqrt{1+\eta^2}} - \frac{\mathbb{E}[E]}{1+\eta^2}\,\Phi^{-1}(\beta)}{\mathbb{E}[E]\,\sigma_\nu/\sqrt{1+\eta^2}} \right) \right.$$
$$\left. \times \frac{(\eta^2)^{k-1}\, e^{-\eta^2/\theta}}{\Gamma(k)\,\theta^k}\, d\eta^2 \right\} \times \frac{\left(\mathbb{E}[E]\right)^{bg} \left(1-\mathbb{E}[E]\right)^{(1-b)g}}{B\left(1+bg,\ 1+(1-b)g\right)}\, d\mathbb{E}[E], \quad (14)$$

where the expression on the first line of the integrand is the normal density implied by Equation 13, and the expression on the second line is the product of the Gamma$(k,\ \theta)$ and Beta$(b,\ g)$ densities. (Within these densities, Γ and B represent the standard gamma and beta functions.)

Using quasi-Monte Carlo methods to evaluate integrals of this sort, we found maximum-likelihood estimates of the parameter vector $(\bar{\nu},\ \sigma_\nu,\ k,\ t,\ b,\ g)$ for the dataset that excludes the outliers discussed in the previous section. These are reported in Table 4. Of course, the raw parameters can be a little difficult to parse directly. To aid in interpretation, in Table 5, we transform the estimates in Table 4 to give a few more intuitive quantities.

Before discussing these though, we first discuss how well the model with literature noise fits. Begin by considering the limit where the parameter g approaches infinity, the parameter t approaches zero, the parameter k is set to $\eta^2/2$, and the parameter b is set to $\mathbb{E}[E]$. In this limit, there is no literature level noise: $\mathcal{E}$ is always equal to a constant $\mathbb{E}[E]$ and H^2 is always equal to a constant η^2. In other words, we recover the no-literature-level-noise model of the previous section.

This means that we can effectively nest our no-literature-level-noise model within our with-literature-level-noise model with two limiting constraints. Hence, we can conduct a likelihood-ratio test to compare our two models: twice the difference in likelihood should be asymptotically distributed χ^2 with two degrees of freedom. This allows us to test the null hypothesis that the with-noise model fits equally as well as the without-noise model. The data overwhelming reject this null; the p-value for this test (0.004) is reported at the bottom of Table 4. Adding the noise leads to a better fit, which suggests that there is indeed significant cross-experiment heterogeneity in threshold width and update strength.

Now, we move on to Table 5. Looking to the upper portion, we learn about where signals tend to lie relative to the prior distribution. The median experiment has a signal that lies around the 63rd percentile of the prior distribution. Such a signal is bad news to about one-third of subjects. Since one standard deviation below the mean signal corresponds to roughly the

Table 4: Maximum-Likelihood Estimates of Raw Structural Parameters

Parameter	Estimate
The signal, N, is distributed $\mathcal{N}(\bar{v}, \sigma_v^2)$.	
$\bar{v}$	0.336
	(0.047)
σ_v	0.259
	(0.046)
The variance of the threshold distribution, H^2, is distributed Gamma$(k,\ t)$.	
k	1.352
	(0.660)
t	10.820
	(8.402)
The expected update strength, $\mathcal{E}$, is distributed Beta$(b,\ g)$.	
b	0.012
	(0.043)
g	20.443
	(66.138)
Papers	21
Studies	71
Model comparison LR test	
With literature-level noise (Section 4.4.b)	
fits better than	
No literature-level noise (Section 4.4.a)	
p-value	0.004

NOTES. These are the maximum-likelihood estimates for the likelihood in Equation 14. The outliers discussed in Section 4.4.a are omitted from the sample. In the limit where $g \to \infty$, $t \to 0$, and $k = \eta^2 / t$, the model with literature-level noise becomes the model without literature-level noise. Since the models are nested, we can conduct a likelihood-ratio test to compare the fits of our two models. Standard errors are in parentheses.

Table 5: Selected Transformations of the Estimates in Table 4

Transformed parameter	Estimate
Key signal realizations as percentiles of the prior distribution	
$\Phi\left(\bar{\nu}-\sigma_{\nu}\right)$	0.531
	(0.028)
$\Phi\left(\bar{\nu}\right)$	0.632
	(0.017)
$\Phi\left(\bar{\nu}+\sigma_{\nu}\right)$	0.724
	(0.020)
Moments of the threshold width, H	
$\mathbb{E}\left[H\right]$	3.494
	(0.663)
$\sqrt{\text{Var}\left[H\right]}$	1.564
	(0.560)
Moments of the expected update strength, $\mathcal{E}$	
$\mathbb{E}\left[\mathcal{E}\right]$	0.055
	(0.109)
$\sqrt{\text{Var}\left[\mathcal{E}\right]}$	0.073
	(0.152)
Papers	21
Studies	71

NOTES. A few notes elucidate where these numbers come from. In the top section, recall that $\Phi(\cdot)$ is the cumulative distribution function for the standard normal. In the middle section, $\mathbb{E}\left[H\right] = \mathbb{E}\left[\sqrt{H^2}\right]$. The (1/2)th moment of a Gamma(k,t) distribution is $\sqrt{t}\ \Gamma(k+1/2)/\Gamma(k)$. In the bottom section, the first and second moments of a Beta(b,g) distribution are $(1+bg)/(2+g)$ and $(1+bg)(2+bg)/\big((2+g)(3+g)\big)$. All standard errors (in parentheses) are computed via the delta method.

50th percentile of the prior distribution, we can conclude that only about 15% of experiments have a signal that is bad news to the median subject. And, since one standard deviation above the mean signal sits around the 72nd percentile of the prior distribution, we can conclude that even in experiments with very high signals, there are still about a quarter of subjects who interpret the signal as bad news.

Moving on to the middle portion, we see that the threshold distribution is about 3.5 times wider than the prior distribution, on average. This makes sense, as thresholds are rooted in preferences, while priors are rooted in information acquisition. Assuming all subjects are getting information from similar sources, it makes sense that priors would be less dispersed. Looking to cross-experiment heterogeneity, holding the threshold distribution width to within one standard deviation around its mean, it can vary from from about 2 to 5 times as wide as the prior distribution. So, there is significant cross-experiment heterogeneity in threshold width, as the likelihood ratio test mentioned above has already shown.

Finally, looking to the bottom section, we see that the cross-experiment mean expected update is 5.5%. Of course, there is a good deal of cross-experiment heterogeneity and noise in the estimates, but it is clear from the numbers that most experiments have an expected update strength well below 25%. Looking back to the theory underpinning this paper, most of it hinges on the expected update strength being small enough for our approximation of the treatment effect to be a good one. The fact that our maximum-likelihood estimates of the expected update strength are indeed small provides an important demonstration of internal validity.

5 Survey Evidence on Current Intuitions

As shown in Section 4, researchers regularly run information experiments in settings with low baseline take-up rates. This is the opposite of what our model prescribes for researchers seeking to maximize treatment effects. Our model suggests that, generally speaking, treatment effects are maximized for high baseline rates. This disparity may suggest that our theory's intuition may not be ex-ante obvious. However, there are many factors that determine where we run experiments. As a result, we cannot be sure if the ideas in our model, even if they are unpublished, are already known by experts.

To answer this question, we measured experts' intuitions of how baseline rates predict treatment effects in information provision experiments. In 2018, we surveyed attendees at the Behavioral Science and Policy Association Annual (BSPA) Conference. Attendees of this conference consist of both academics and policymakers interested in interventions for behavioral change.[55] Our survey asked participants to read a scenario and answer

55. See https://behavioralpolicy.org/bspa-events/bspa-annual-conference-2018/ for de-

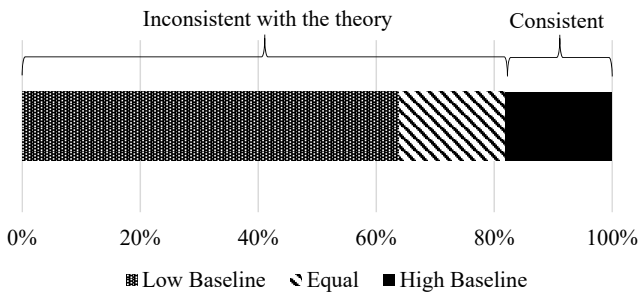


Figure 6: Expert Beliefs: Would You Expect Higher Treatment Effect with Low or High Baseline Rate?

three questions.[56]

The hypothetical scenario was about a policymaker who wants to run an information-provision intervention to increase the number of women who get breast cancer screenings (a binary decision for each woman). The policymaker only has enough money to run the intervention in one of two sites. The two sites are identical other than the baseline rate of women who already get mammograms. Her only objective is to increase the number of mammograms received (i.e., the magnitude of the treatment effect). The first question asks about the site at which the policymaker should run the information intervention. The second question is a free response that asks why the subject gave their answer to the first question. Third, the survey asks whether the subject is a policymaker or an academic and, if they are an academic, their field. The survey was done in private, with paper and pen, and placed into an envelope to ensure anonymity of responses. The full text of the survey distributed at the BSPA can be found in Appendix B.

A large majority of responses are inconsistent with the intuition of our model. 72 subjects filled out our survey. Figure 6 shows responses to the first question. Only 13 of 72 (18%) said the policymaker should choose the high baseline site, the answer consistent with our theory. 64% said the low baseline site would be optimal, and 18% believe the baseline is immaterial to the treatment effect. No subgroup of respondents, not psychologists, economists, or policymakers, provide answers consistent with our model with any regularity.

Further, no respondent provides an explanation for their answer that is in line with our theory. The modal explanation when a subject chooses the high baseline site is that people like doing what a majority of other people

tails.

56. Potential participants were recruited by a research assistant during a coffee break between conference sessions. Potential participants were only told it was a "survey" that would take "a few minutes". Only clarifying questions were answered. The research assistant did not know the correct answer at the time of the survey. Participants filled out the survey individually.

are doing.[57] Meanwhile, the modal explanation when the low baseline site is chosen is, as one respondent succinctly put it, "There's more women to 'switch' from 'no' to 'yes'."

Though Section 4 suggests our model is first order in predicting the outcome of an information experiment, the model's intuition has not been a part of our collective intuition. Very few policymakers and academics specifically interested in behavioral-change interventions give answers in line with our model, and zero provide an explanation related to the theory's intuition.

6 Practitioner's Guide

The results from the theory and meta-analysis provide a number of potentially useful insights for practitioners who may be considering using information interventions to nudge people deciding between taking an action or not. This section aims to summarize these insights.

First, our paper highlights the importance of attempting to identify what baseline take-up will be in the absence of an information intervention. In many settings, previous choices of similar agents can provide a benchmark.[58] Armed with this knowledge, a practitioner can use our results (highlighted in the next two insights) to assess whether using an information intervention will improve take-up.

Second, as described in the previous sections, low baseline take-up environments are more likely to generate backfires.[59] If a potential nudge is similar to those that have been tested in the literature (with regard to how their information content compares to agents' prior beliefs), we estimate that a baseline of 0.10 or below is likely to generate a backfire. To decrease the likelihood of a backfire below 10%, we estimate the baseline rate should be roughly higher than 0.50.

Third, settings in which many agents are expected to take-up (i.e., settings with a high baseline) may be particularly ripe for information nudges to have big positive impacts. Given our parameter estimates, the treatment effect of the "typical" nudge is expected to have the largest positive treatment effect at a baseline of 0.75.

57. These answers suggest some respondents might assume the information provided would be social information (e.g., "X% of other women get mammograms", though that was not mentioned in the scenario). As a result, the number of people choosing the answer consistent with our model might be an overestimate.

58. In Coffman, Featherstone, and Kessler (2017), we found the rate of accepting a job offer at Teach For America is relatively stable year to year. Available data suggests the same is true in many of the other outcomes analyzed by the experiments in our meta-analysis.

59. As mentioned previously, the baseline rate is the incidence rate of taking the desired action absent the information that will be provided. This means that if the information is provided through an email campaign, the correct baseline rate is from a control group who would receive an email but not the information.

These latter two results run counter to the standard intuitions revealed by practitioners—they ran counter to our intuitions before writing this paper—suggesting they may be the most important insights to heed. Importantly, the failure of information interventions to encourage take-up of an action in settings with low baselines does not imply that no nudge will be effective there. It just indicates that information nudges are unlikely to be helpful at encouraging a desired behavior. Consequently, practitioners may want to choose a different tool from their toolbox when in those settings. Similarly, practitioners may want to think seriously about using information nudges in settings with high baselines. While we did not observe many experiments with high baselines in the meta-analysis, our theory and empirical results suggest that those settings may the perfect targets for such information nudges.

Fourth, consistent with standard intuition, our model also predicts that nudges that provide better news (i.e., nudges that provide information that is higher in the agents' prior belief distribution) will be less likely to have a negative treatment effect (and any negative treatment effects will likely be smaller in magnitude) and be more likely to have a positive treatment effect (and any positive treatment effects will likely be larger in magnitude). This means that a nudge that is exceedingly high in agents' prior belief distribution can be effective, even at low baselines. However, note that for a nudge that provides truthful information to be that high in agents' prior belief distribution suggests the set of agents is incredibly pessimistic. Such environments may be unlikely to arise in practice. Our structural analysis estimated that the 95% confidence interval for the strength of the average nudge in our data spanned from the 40th to the 95th percentile of agents' priors, suggesting that nudges that are good news to everyone may be few and far between.

Fifth, the previous result highlights the value of collecting belief information from agents. While rarely done in practice, surveying agents about their beliefs to assess where a particular nudge falls in the distribution of agents' priors would give the practitioner additional information about the likelihood of success of that nudge.

Finally, our model can provide structure for considering results from previous information interventions (e.g., all those performed by a "nudge unit"). When faced with many null results or backfires, it could be natural to assume information interventions do not work . But if these negative and null effects are occurring at low baselines, our model would suggest that the information may be working exactly how we would expect.

7 Conclusion

As nudges become more prominent in the academic literature and more common as a policy tool, there is an enhanced interest in understanding

why nudges work and when—or for whom—they will be successful (see, e.g., Beshears, Choi, Laibson, Madrian, and Wang 2015). Central to this exercise is developing models of these nudges that can give insight into the underlying mechanisms. In this paper, we introduce a theory of information nudges that allows for Bayesian updating in a setting of binary choice.

Our model highlights that in these settings, the relevant question about the sign and magnitude of the treatment effect is whether the information nudge provides good news about taking the action to agents *at the margin*. Our model additionally suggests that baseline take-up rate in the untreated group can be a useful proxy for the beliefs of marginal agents. This allows researchers and practitioners to infer the likely sign and magnitude of a treatment effect arising from an information nudge even without information on beliefs. In a meta-analysis of information experiments, we find that the relationship between treatment effect size and baseline take-up rate matched the pattern predicted by the theory, allowing us to rationalize previously puzzling results from the literature.

Both the reduced-form and the structural meta-analysis provide insights for practitioners. First, information nudges may backfire on populations with low baseline take-up. Second, the positive effect of an information intervention is maximized for baseline take-up around 0.75. Given that the median baseline in the experiments we found for our meta-analysis is 0.34, and given that about a third of baselines are below 0.23, these lessons do not appear to have entered our collective wisdom yet.

Though our meta-analysis was sharply focused on a specific type of information intervention, other nudges may work partially through information channels. For example, reminders, which are often assumed to work through inattention (e.g., Taubinsky 2014), have been shown to affect beliefs about the probability others take a certain action (see, e.g., Del Carpio 2013) and so might also work through an information channel. To the extent that information is active, the main insights of our model would still apply. We hope future work models other impactful nudges to understand when they are helpful, how to maximize their efficacy, and when they may backfire.[60] Modeling these nudges can unleash their full potential.

References

Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9): 1082–1095.

60. For example, nudges in choice architecture are a potentially powerful policy tool; however, they do not always deliver the expected results (see, e.g., Kessler and Roth 2015 on "yes–no" vs. "opt-in" choice frames for active choice organ donation requests of the type made at Departments of Motor Vehicles).

Allcott, Hunt, and Judd B Kessler. 2019. "The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons." *American Economic Journal: Applied Economics* 11 (1): 236–76.

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105 (8): 2501–38.

Avitabile, Ciro, and Rafael de Hoyos. 2018. "The Heterogeneous Effect of Information on Student Performance: Evidence From a Randomized Control Trial in Mexico." *Journal of Development Economics* 135:318–348.

Beaton, Albert E., and John W. Tukey. 1974. "The Fitting of Power Series, Meaning Polynomials, Illustrated on Band-Spectroscopic Data." *Technometrics* 16 (2): 147–185.

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 2005. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons.

Bernheim, B. Douglas, Andrey Fradkin, and Igor Popov. 2015. "The Welfare Economics of Default Options in 401(k) Plans." *American Economic Review* 105 (9): 2798–2837.

Beshears, John, James J. Choi, David Laibson, Brigitte C. Madrian, and Katherine L. Milkman. 2015. "The Effect of Providing Peer Information on Retirement Savings Decisions." *The Journal of Finance* 70 (3): 1161–1201.

Beshears, John, James J. Choi, David Laibson, Brigitte C. Madrian, and Sean (Yixiang) Wang. 2015. "Who Is Easier to Nudge?" *Working paper*.

Bhargava, Saurabh, and Dayanand Manoli. 2013. "Why Are Benefits Left on the Table? Assessing the Role of Information, Complexity, and Stigma on Take-Up with an IRS Field Experiment." *American Economic Review*.

Bollinger, Bryan, Phillip Leslie, and Alan Sorensen. 2011. "Calorie Posting in Chain Restaurants." *American Economic Journal: Economic Policy* 3 (1): 91–128.

Breusch, T.S., and A.R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* 47 (5): 1287–1294.

Brown, Martin, Stefan T Trautmann, and Razvan Vlahu. 2017. "Understanding bank-run contagion." *Management Science* 63 (7): 2272–2282.

Butera, Luigi, Robert Metcalfe, William Morrison, and Dmitry Taubinsky. 2022. "Measuring the welfare effects of shame and pride." *American Economic Review* 112 (1): 122–168.

Cai, Hongbin, Yuyu Chen, and Hanming Fang. 2009. "Observational Learning: Evidence from a Randomized Natural Field Experiment." *American Economic Review* 99 (3): 864.

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez. 2012. "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction." *American Economic Review* 102 (6): 2981–3003.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *Quarterly Journal of Economics* 124 (4): 1639–1674.

Chambers, Christopher P., and Paul J. Healy. 2012. "Updating Toward the Signal." *Economic Theory* 50 (3): 765–786.

Chen, Yan, F. Maxwell Harper, Joseph Konstan, and Sherry Xin Li. 2010. "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens." *American Economic Review,* 1358–1398.

Chung, Kai Lai. 2001. *A Course In Probability Theory*. Academic Press.

Cialdini, Robert B., Linda J. Demaine, Brad J. Sagarin, Daniel W. Barrett, Kelton Rhoads, and Patricia L. Winter. 2006. "Managing Social Norms for Persuasive Impact." *Social Influence* 1 (1): 3–15.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren. 1990. "A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places." *Journal of Personality and Social Psychology* 58 (6): 1015.

Clark, Charles E. 1962. "The PERT Model for the Distribution of an Activity Time." *Operations Research* 10 (3).

Clark, Robert L., Jennifer A. Maki, and Melinda Sandler Morrill. 2014. "Can Simple Informational Nudges Increase Employee Participation in a 401(k) Plan?" *Southern Economic Journal* 80 (3): 677–701.

Coffman, Lucas C., Clayton R. Featherstone, and Judd B. Kessler. 2017. "Can Social Information Affect What Job You Choose and Keep?" *American Economic Journal: Applied Economics* 9 (1): 96–117.

Cook, R. Dennis. 1977. "Detection of Influential Observation in Linear Regression." *Technometrics* 42 (1): 65–68.

Croson, Rachel, and Jen Yue Shang. 2008. "The Impact of Downward Social Information on Contribution Decisions." *Experimental Economics* 11 (3): 221–233.

Del Carpio, Lucia. 2013. "Are the Neighbors Cheating? Evidence from a Social Norm Experiment on Property Taxes in Peru." *Working paper*.

DellaVigna, Stefano, and Elizabeth Linos. 2022. "RCTs to scale: Comprehensive evidence from two nudge units." *Econometrica* 90 (1): 81–116.

Diaconis, Persi, and Donald Ylvisaker. 1979. "Conjugate Priors for Exponential Families." *The Annals of statistics* 7 (2): 269–281.

Fellner, Gerlinde, Rupert Sausgruber, and Christian Traxler. 2013. "Testing Enforcement Strategies in the Field: Threat, Moral Appeal and Social Information." *Journal of the European Economic Association* 11 (3): 634–660.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." *Economics Letters* 71 (3): 397–404.

Frey, Bruno S., and Stephan Meier. 2004. "Social Comparisons and Pro-Social Behavior: Testing "Conditional Cooperation" in a Field Experiment." *American Economic Review,* 1717–1722.

Gerber, Alan S., and Todd Rogers. 2009. "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and So Should You." *Journal of Politics* 71 (01): 178–191.

Goldstein, Noah J., Robert B. Cialdini, and Vladas Griskevicius. 2008. "A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels." *Journal of Consumer Research* 35 (3): 472–482.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart. 2023. "Designing information provision experiments." *Journal of economic literature* 61 (1): 3–40.

Hallsworth, Michael, John List, Robert Metcalfe, and Ivo Vlaev. 2014. *The Behavioralist as Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance*. Technical report. National Bureau of Economic Research.

Hastings, Justine, Christopher A. Neilson, and Seth D. Zimmerman. 2015. *The Effects of Earnings Disclosure on College Enrollment Decisions*. Technical report. National Bureau of Economic Research.

Hastings, Justine S., and Jeffrey M. Weinstein. 2007. *Information, School Choice, and Academic Achievement: Evidence from Two Experiments*. Technical report. National Bureau of Economic Research.

Huber, Peter J. 1964. "Robust Estimation of a Location Parameter." *The Annals of Mathematical Statistics,* 73–101.

Huber, Peter J., and Elvezio M. Ronchetti. 2011. *Robust Statistics*. John Wiley & Sons.

Jensen, Robert. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *Quarterly Journal of Economics* 125 (2): 515–548.

Jessoe, Katrina, and David Rapson. 2014. "Knowledge Is (Less) Power: Experimental Evidence from Residential Energy Use." *American Economic Review* 104 (4): 1417–38.

Karadja, Mounir, Johanna Mollerstrom, and David Seim. 2017. "Richer (and Holier) Than Thou? The Effect of Relative Income Improvements on Demand for Redistribution." *Review of Economics and Statistics* 99 (2): 201–212.

Keser, Claudia, and Frans Van Winden. 2000. "Conditional Cooperation and Voluntary Contributions to Public Goods." *Scandinavian Journal of Economics* 102 (1): 23–39.

Kessler, Judd B., and Alvin E. Roth. 2015. "Organ Allocation Policy and the Decision to Donate." *American Economic Review*.

Koenker, Roger. 1981. "A Note on Studentizing a Test for Heteroscedasticity." *Journal of Econometrics* 17 (1): 107–112.

Kuziemko, Ilyana, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva. 2015. "How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments." *American Economic Review* 105 (4): 1478–1508.

Li, Guoying. 1985. "Robust regression." *Exploring Data Tables, Trends, and Shapes* 281:U340.

Lo, Ambrose. 2019. "Demystifying the Integrated Tail Probability Expectation Formula." *The American Statistician* 73 (4): 367–374.

Martin, Richard, and John Randal. 2008. "How is Donation Behaviour Affected by the Donations of Others?" *Journal of Economic Behavior & Organization* 67 (1): 228–238.

Nguyen, Trang. 2008. "Information, Role Models and Perceived Returns to Education: Experimental Evidence from Madagascar." *Working paper*.

Owen, Donald Bruce. 1980. "A Table of Normal Integrals." *Communications in Statistics—Simulation and Computation* 9 (4): 389–419.

Pomeranz, Dina. 2015. "Taxation without information." *American Economic Review* 105 (8): 2539–69.

Potters, Jan, Martin Sefton, and Lise Vesterlund. 2005. "After You—Endogenous Sequencing in Voluntary Contribution Games." *Journal of Public Economics* 89 (8): 1399–1419.

Salganik, Matthew J., Peter Sheridan Dodds, and Duncan J. Watts. 2006. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market." *Science* 311 (5762): 854–856.

Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2007. "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science* 18 (5): 429–434.

Shang, Jen, and Rachel Croson. 2009. "A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods." *Economic Journal* 119 (540): 1422–1439.

Simonsohn, Uri. 2018. "Two Lines: A Valid Alternative to the Invalid Testing of U-Shaped Relationships with Quadratic Regressions." *Advances in Methods and Practices in Psychological Science* 1 (4): 538–555.

Sunstein, Cass R., and Richard H. Thaler. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Penguin Books.

Taubinsky, Dmitry. 2014. "From Intentions to Actions: A Model and Experimental Evidence of Inattentive Choice." *Working Paper*.

Vesterlund, Lise. 2003. "The Informational Value of Sequential Fundraising." *Journal of Public Economics* 87 (3): 627–657.

Whitehead, Mark, Rhys Jones, Rachel Howell, Rachel Lilley, and Jessica Pykett. 2014. *Nudging All Over the World: Assessing the Impacts of the Behavioural Sciences on Public Policy*. Report. Economic and Social Research Council (United Kingdom).

Wooldridge, J.M. 2019. *Introductory Econometrics: A Modern Approach*. 7th ed. Cengage.

APPENDIX

A Theory of Information Nudges

Lucas C. Coffman
Boston College

Clayton R. Featherstone
Baylor University

Judd B. Kessler
The Wharton School, University of Pennsylvania

A Meta-Analysis Appendix

A.1 Selection of papers

At the end of this section of the appendix, we have included two large-form figures. The first is a copy of the email we used to elicit suggested papers for inclusion in our meta-analysis. The second is a table that goes through our rationale for either including or excluding each paper we recieved.

A.2 Reduced-Form Meta-Analysis

To see all estimates of the three line analysis, with other break points, we have included Table A.1 at the end of this section of the appendix.

A.3 The model of Section 4.4.a: no literature-level noise

A.3.a Theoretical derivation of the distribution of τ

Here, we restate and prove the result from Section 4.4.a.

Result. *For the parametric model of this section, when the realization of (N, B) is (ν, β), the treatment effect is given by*

$$\tau_{\beta,\nu} = \varphi\left(\Phi^{-1}(\beta)\right)\left\{\frac{\mathbb{E}[E]}{\sqrt{1+\eta^2}}\,\nu \;+\; \frac{\mathbb{E}[E]}{1+\eta^2}\,\Phi^{-1}(\beta)\right\}.$$

Since N is distributed normally with mean $\bar{\nu}$ and variance σ_ν^2, this means that, conditional on the baseline, B, having realization β, the treatment effect is normally distributed, with mean and standard deviation given by

$$\mathbb{E}\left[\tau_{B,N}\,\middle|\,B=\beta\right] = \varphi\left(\Phi^{-1}(\beta)\right)\left\{\frac{\mathbb{E}[E]}{\sqrt{1+\eta^2}}\,\bar{\nu} \;+\; \frac{\mathbb{E}[E]}{1+\eta^2}\,\Phi^{-1}(\beta)\right\},$$

$$\sqrt{\mathrm{Var}\left[\tau_{B,N}\,\middle|\,B=\beta\right]} = \varphi\left(\Phi^{-1}(\beta)\right)\frac{\mathbb{E}[E]\,\sigma_\nu}{\sqrt{1+\eta^2}}. \tag{A.15}$$

From this, it almost immediately follows that the conditional-on-baseline probability of a positive treatment effect is given by

$$\Pr\left\{\tau_{B,N} > 0 \;\middle|\; B=\beta\right\} \;=\; \Phi\left(\frac{\mathbb{E}\left[\tau_{B,N}\,\middle|\,B=\beta\right]}{\sqrt{\mathrm{Var}\left[\tau_{B,N}\,\middle|\,B=\beta\right]}}\right). \tag{A.16}$$

Proof. Integrals 110 and 111 in Owen (1980) are equivalent to

$$\int_{-\infty}^{\infty} \frac{1}{b}\varphi\left(\frac{x-a}{b}\right)\varphi(x)\,dx \;=\; \frac{1}{\sqrt{1+b^2}}\,\varphi\left(-\frac{a}{\sqrt{1+b^2}}\right) \quad \text{and}$$

$$\int_{-\infty}^{\infty} x\,\frac{1}{b}\varphi\left(\frac{x-a}{b}\right)\varphi(x)\,dx \;=\; \frac{a}{1+b^2}\,\frac{1}{\sqrt{1+b^2}}\,\varphi\left(-\frac{a}{\sqrt{1+b^2}}\right).$$

(These integrals can also be derived by the completing the square in the exponents of the integrands.) Using the first integral (with $a = -\Phi^{-1}(\beta)\sqrt{1+\eta^2}$ and $b = \eta$), we can compute

$$\begin{aligned} f^{\Delta}(0) &= \int_{-\infty}^{\infty} f^{\Theta}(\mu)\,\varphi(\mu)\,d\mu \\ &= \int_{-\infty}^{\infty} \frac{1}{\eta}\,\varphi\left(\frac{\mu + \Phi^{-1}(\beta)\sqrt{1+\eta^2}}{\eta}\right)\varphi(\mu)\,d\mu = \frac{1}{\sqrt{1+\eta^2}}\,\varphi\left(\Phi^{-1}(\beta)\right). \end{aligned}$$

Using the second integral, we can compute

$$\begin{aligned} \mathbb{E}\left[M \,\middle|\, \Delta = 0\right] &= \frac{1}{f^{\Delta}(0)} \int_{-\infty}^{\infty} \mu\, f^{\Theta}(\mu)\,\varphi(\mu)\,d\mu \\ &= \frac{1}{f^{\Delta}(0)} \int_{-\infty}^{\infty} \mu\,\frac{1}{\eta}\,\varphi\left(\frac{\mu + \Phi^{-1}(\beta)\sqrt{1+\eta^2}}{\eta}\right)\varphi(\mu)\,d\mu = -\frac{\Phi^{-1}(\beta)}{\sqrt{1+\eta^2}}. \end{aligned}$$

The treatment effect is given by

$$\mathbb{E}\left[\tau_{B,N} \,\middle|\, B = \beta\right] = \mathbb{E}\left[E\right]\, f_{\beta}^{\Delta}(0)\,\left\{\nu - \mathbb{E}_{\beta}\left[M \,\middle|\, \Delta = 0\right]\right\}.$$

Plugging in our integrals from above immediately yields the treatment effect, $\tau_{\beta,\nu}$, in terms of the signal realization, ν.

Since ν is $\mathcal{N}(\bar{\nu}, \sigma_{\nu}^2)$ and τ is a linear function of ν, we know τ must be normally distributed. The expressions for $\mathbb{E}\left[\tau_{B,N} \,\middle|\, B = \beta\right]$ and $\mathrm{Var}\left[\tau_{B,N} \,\middle|\, B = \beta\right]$ follow from the well-known facts that $\mathbb{E}\left[a + bX\right] = a + b\mathbb{E}\left[X\right]$ and $\mathrm{Var}\left[a + bX\right] = b^2\,\mathrm{Var}\left[X\right]$.

Finally, given the previous paragraph, the conditional-on-baseline probability of a positive treatment effect can be written

$$\Pr\left\{\tau_{B,N} > 0 \;\middle|\; B = \beta\right\} \;=\; 1 - \Phi\left(-\frac{\mathbb{E}\left[\tau_{B,N} \,\middle|\, B = \beta\right]}{\sqrt{\mathrm{Var}\left[\tau_{B,N} \,\middle|\, B = \beta\right]}}\right).$$

The symmetry of the normal distribution shows this is equivalent to the expression in the statement of the result. □

A.3.b Technical analysis of outliers

To diagnose which studies are causing the problem, we consider an approach inspired by Belsley, Kuh, and Welsch (2005), looking at the leverage that each data point has on parameter estimates. For each study in our sample, we estimate the regression without it and compute how far such leave-one-out estimates are from the full-sample estimates, normalizing by the full-sample estimates' standard errors. Using the parameter a as an example then, our leverage measure for study i in the sample is

$$\ell_i(a) \equiv \frac{\left|\hat{a} - \hat{a}^{(-i)}\right|}{\hat{\sigma}_a},$$

where $\hat{a}$ and $\hat{\sigma}_a$ are the full-sample estimate of a and its standard error, and $\hat{a}^{(-i)}$ is the estimate of a when study i is left out from the sample. The

leverage measures $\ell_i(b)$ and $\ell_i(\sigma_\zeta)$ are defined analogously. From there, we define each observation's leverage, ℓ_i, to be the maximum of $\ell_i(a)$, $\ell_i(b)$, and $\ell_i(\sigma_\zeta)$. For our regression, the four largest values of ℓ_i are 1.67, 1.14, 1.06, and 0.98. The next highest value is 0.37. The mean of the four largest ℓ_i is 13 times larger than the mean of the rest. On the basis of these leverage numbers, we will treat the four studies with the largest ℓ_i values as outliers. Essentially, if a study's inclusion moves a parameter estimate by more than about one standard error, we exclude it. Note that this more sophisticated method chooses the same outliers as the visual-inspection method discussed in the main text.

Once we have identified the outliers, it is worth testing the no-literature-level-noise model's assumption of homoskedasticity in the regression of Equation 13. In the full sample, if we run a Breusch–Pagan test (Breusch and Pagan 1979; Koenker 1981; Wooldridge 2019) on our full-sample regression, the null hypothesis of homoskedasticity is overwhelmingly rejected ($p = 0.003$). If we remove the outliers, the same test fails to reject the null at any conventional level of significance ($p = 0.51$).

A.4 Long-format figures and tables

Table A.1: Three-line regressions
Testing down-up-down relationship

Break points					
Lower	Upper	β_1	β_2	β_3	$RSS^\dagger$
0.07	0.50	-0.371	0.153	-0.053	0.243
	s.e.:	(0.339)	(0.057)	(0.136)	
	p-value:	*0.00*	*0.01*	*0.41*	

Notes: Reports OLS estimates of three-line fits of the data, for various break points indicated in first two columns. Each set of two rows is a new regression, reporting the estimates first, then the *p*-values. β_1 is the estimate for the first line, β_2 for the second, and β_3 for the third. Standard errors are the standard deviation of estimates from hierarchical bootstrap described in Section 4.3.a modified as described in Section 4.3.b, footnote 47. Note the boot-strapped distributions of estimates are skewed, left-skewed for (I) and (III), and right-skewed for (II). "*p*-value" reports the percentage of bootstrapped estimates that are inconsistent with the model's prediction: i.e. non-negative for (I) and (III) and non-positive for (II). †Residual sum of squares is the average RSS from the bootstrap procedure described in Section 4.3.b. When values are equal to a break point, and are thus used to estimate the line below and the line above, the RSS is adjusted such that, for each datum on a break point, half of its squared residual for each the lower line and the upper line is added to the sum.

continued on next page

Table A.1 – continued from previous page

Break points Lower	Upper	β_1	β_2	β_3	$RSS^\dagger$
0.09	0.50	-0.167	0.204	-0.053	0.241
	s.e.:	(0.066)	(0.096)	(.136)	
	p-value:	*0.04*	*0.03*	*0.41*	
0.12	0.50	-0.242	0.197	-0.053	0.241
	s.e.:	(0.101)	(0.162)	(0.136)	
	p-value:	*0.05*	*0.16*	*0.41*	
0.23	0.50	0.502	0.392	-0.053	0.247
	s.e.:	(0.245)	(0.207)	(0.136)	
	p-value:	*0.94*	*0.28*	*0.41*	
0.07	0.60	-0.371	0.124	-0.656	0.241
	s.e.:	(0.339)	(0.044)	(0.40)	
	p-value:	*0.00*	*0.00*	*0.20*	
0.09	0.60	-0.167	0.145	-0.656	0.240
	s.e.:	(0.066)	(0.066)	(0.40)	
	p-value:	*0.04*	*0.02*	*0.20*	
0.12	0.60	-0.242	0.116	-0.656	0.239
	s.e.:	(0.101)	(0.095)	(0.40)	
	p-value:	*0.05*	*0.14*	*0.20*	
0.23	0.60	0.502	0.210	-0.656	0.245
	s.e.:	(0.245)	(0.111)	(0.40)	
	p-value:	*0.94*	*0.25*	*0.20*	
0.07	0.70	-0.371	0.162	0.360	0.239

Notes: Reports OLS estimates of three-line fits of the data, for various break points indicated in first two columns. Each set of two rows is a new regression, reporting the estimates first, then the *p*-values. β_1 is the estimate for the first line, β_2 for the second, and β_3 for the third. Standard errors are the standard deviation of estimates from hierarchical bootstrap described in Section 4.3.a modified as described in Section 4.3.b, footnote 47. Note the boot-strapped distributions of estimates are skewed, left-skewed for (I) and (III), and right-skewed for (II). "*p*-value" reports the percentage of bootstrapped estimates that are inconsistent with the model's prediction: i.e. non-negative for (I) and (III) and non-positive for (II). †Residual sum of squares is the average RSS from the bootstrap procedure described in Section 4.3.b. When values are equal to a break point, and are thus used to estimate the line below and the line above, the RSS is adjusted such that, for each datum on a break point, half of its squared residual for each the lower line and the upper line is added to the sum.

continued on next page

Table A.1 – continued from previous page

Break points					
Lower	Upper	β_1	β_2	β_3	$RSS^\dagger$
	s.e.:	(0.339)	(0.033)	(0.541)	
	p-value:	*0.00*	*0.00*	*0.58*	
0.09	0.70	-0.167	0.189	0.360	0.237
	s.e.:	(0.066)	(0.046)	(0.541)	
	p-value:	*0.04*	*0.00*	*0.58*	
0.12	0.70	-0.242	0.184	0.360	0.238
	s.e.:	(0.101)	(0.061)	(0.541)	
	p-value:	*0.05*	*0.01*	*0.58*	
0.23	0.70	0.502	0.245	0.360	0.243
	s.e.:	(0.245)	(0.068)	(0.541)	
	p-value:	*0.94*	*0.02*	*0.58*	
0.07	0.73	-0.371	0.121	0.037	0.257
	s.e.:	(0.339)	(0.026)	(1.850)	
	p-value:	*0.04*	*0.00*	*0.49*	
0.09	0.73	-0.167	0.130	0.037	0.256
	s.e.:	(0.066)	(0.034)	(1.850)	
	p-value:	*0.04*	*0.00*	*0.49*	
0.12	0.73	-0.242	0.116	0.037	0.256
	s.e.:	(0.101)	(0.042)	(1.850)	
	p-value:	*0.05*	*0.01*	*0.49*	
0.23	0.73	0.502	0.155	0.037	0.262
	s.e.:	(0.245)	(0.045)	(1.850)	

Notes: Reports OLS estimates of three-line fits of the data, for various break points indicated in first two columns. Each set of two rows is a new regression, reporting the estimates first, then the *p*-values. β_1 is the estimate for the first line, β_2 for the second, and β_3 for the third. Standard errors are the standard deviation of estimates from hierarchical bootstrap described in Section 4.3.a modified as described in Section 4.3.b, footnote 47. Note the boot-strapped distributions of estimates are skewed, left-skewed for (I) and (III), and right-skewed for (II). "*p*-value" reports the percentage of bootstrapped estimates that are inconsistent with the model's prediction: i.e. non-negative for (I) and (III) and non-positive for (II). †Residual sum of squares is the average RSS from the bootstrap procedure described in Section 4.3.b. When values are equal to a break point, and are thus used to estimate the line below and the line above, the RSS is adjusted such that, for each datum on a break point, half of its squared residual for each the lower line and the upper line is added to the sum.

continued on next page

Table A.1 – continued from previous page

Break points					
Lower	Upper	β_1	β_2	β_3	$RSS^\dagger$
	p-value:	*0.94*	*0.02*	*0.49*	

Notes: Reports OLS estimates of three-line fits of the data, for various break points indicated in first two columns. Each set of two rows is a new regression, reporting the estimates first, then the p-values. β_1 is the estimate for the first line, β_2 for the second, and β_3 for the third. Standard errors are the standard deviation of estimates from hierarchical bootstrap described in Section 4.3.a modified as described in Section 4.3.b, footnote 47. Note the boot-strapped distributions of estimates are skewed, left-skewed for (I) and (III), and right-skewed for (II). "p-value" reports the percentage of bootstrapped estimates that are inconsistent with the model's prediction: i.e. non-negative for (I) and (III) and non-positive for (II). †Residual sum of squares is the average RSS from the bootstrap procedure described in Section 4.3.b. When values are equal to a break point, and are thus used to estimate the line below and the line above, the RSS is adjusted such that, for each datum on a break point, half of its squared residual for each the lower line and the upper line is added to the sum.

Information Experiments (including file drawer)

Judd Kessler <judd.kessler@wharton.upenn.edu> Mon, Dec 7, 2015 at 12:38 PM
To: esa-discuss@googlegroups.com

Hello all,

Lucas Coffman, Clayton Featherstone and I are looking for experimental papers from the lab or field that meet the following two criteria:

1. At least one experimental treatment is attempting to influence a binary action by providing truthful information to subjects (e.g. telling subjects the % of others who take an action, telling subjects the benefits of taking a certain action, etc.).

2. The paper reports the rate of taking the binary action in the control group, the treatment effect, and the standard error of the treatment effect (or these can be imputed).

In particular, we are looking for papers that provide information as a "nudge" that might lead subjects to update their beliefs. (Note: we are not looking for herding or cascade experiments.) The paper does not need to be published — in fact, we are eager to see unpublished manuscripts and manuscripts with null results. If you are unsure about whether to send a paper, please do so.

Papers we have already identified are listed below.

Thanks in advance,

Lucas Coffman, Clayton Featherstone and Judd Kessler

Papers already identified:

Allcott, Hunt and Dmitry Taubinsky. 2015. "Evaluating Behaviorally-Motivated Policy: Experimental Evidence from the Lightbulb Market." American Economic Review, 105(8):2501-2538.

Avitabile, Ciro, and Rafael E. De Hoyos Navarro. 2015. "The Heterogeneous Effect of Information on Student Performance: Evidence from a Randomized Control Trial in Mexico." Working Paper.

Beshears, John, James J. Choi, David Laibson, Brigitte C. Madrian, and Katherine L. Milkman. 2015. "The effect of providing peer information on retirement savings decisions." The Journal of Finance, 70(3): 1161-1201.

Bettinger, Eric P, Bridget Terry Long, Philip Oreopoulos, and Lisa Sanbonmatsu. 2012. "The role of application assistance and information in college decisions: Results from the H&R Block fafsa experiment." The Quarterly Journal of Economics, 127(3): 1205–1242.

Bhargava, Saurabh, and Dayanand Manoli. 2015. "Psychological Frictions and the Incomplete Take-Up of Social Benefits: Evidence from an IRS Field Experiment." American Economic Review, 105(11): 3489-3529.

Cai, Hongbin, Yuyu Chen, and Hanming Fan. 2009. "Observational Learning: Evidence from a Randomized Natural Field Experiment." American Economic Review, 99(3): 864-82.

Clark, Robert L., Jennifer A. Maki, and Melinda Sandler Morrill. 2014. "Can Simple Informational Nudges Increase Employee Participation in a 401 (k) Plan?" Southern Economic Journal, 80(3): 677-701.

Cialdini, Robert, Linda Demaine, Brad Sagarin, Daniel Barrett, Kelton Rhoads, and Patricia Winter. 2006. "Managing Social Norms for Persuasive Impact." Social Influence, 1(1): 3-15.

Coffman, Lucas C., Clayton R. Featherstone, and Judd B. Kessler. 2014. "Can Social Information Affect What Job You Choose and Keep?" Working Paper.

Del Carpio, Lucia. 2014. "Are the Neighbors Cheating? Evidence From a Social Norm Experiment on Property Taxes in Peru." Working Paper.

Fellner, Gerlinde, Rupert Sausgruber, and Christian Traxler. 2013. “Testing enforcement strategies in the field: Threat, moral appeal and social information.” Journal of the European Economic Association, 11(3): 634-660.

Frey, Bruno S, and Stephan Meier. 2004. “Social comparisons and pro-social behavior: Testing “conditional cooperation” in a field experiment.” American Economic Review, 94 (5): 1717–1722.

Goldstein, Noah J, Robert B Cialdini, and Vladas Griskevicius. 2008. “A room with a viewpoint: Using social norms to motivate environmental conservation in hotels.” Journal of Consumer Research, 35(3): 472–482.

Hallsworth, Michael, John A. List, Robert D. Metcalfe, Ivo Vlaev. 2014. “The Behavioralist As Tax Collector: Using Natural Field Experiments to Enhance Tax Compliance.” Working Paper.

Hastings, Justine S., and Jeffrey M. Weinstein. 2008. "Information, School Choice, and Academic Achievement: Evidence from Two Experiments." The Quarterly Journal of Economics 123(4): 1373-1414.

Hastings, Justine, Christopher A. Neilson, and Seth D. Zimmerman. 2015. “The effects of earnings disclosure on college enrollment decisions.” Working Paper.

Jensen, Robert. 2010. “The (perceived) returns to education and the demand for schooling.” The Quarterly Journal of Economics, 125(2): 515–548.

Krupka, Erin and Roberto A. Weber. 2009. “The Focusing and Informational Effects of Norms on Pro-Social Behavior.” Journal of Economic Psychology, 30: 307-320.

Liebman, Jeffrey B., and Erzo FP Luttmer. 2015. “Would People Behave Differently If They Better Understood Social Security? Evidence from a Field Experiment.” American Economic Journal: Economic Policy, 7(1): 275-99.

--
Judd Benjamin Kessler
Assistant Professor, Department of Business Economics and Public Policy
The Wharton School at the University of Pennsylvania
http://assets.wharton.upenn.edu/~juddk/

Paper	Included	Reason for Exclusion	Further Explanation	Source
Allcott, H. and Taubinsky, D., 2015. Evaluating behaviorally motivated policy: experimental evidence from the lightbulb market. *The American Economic Review*, 105(8), pp.2501-2538.	Yes			Initial Search
Andreoni, J. and Mylovanov, T., 2012. Diverging opinions. *American Economic Journal: Microeconomics*, 4(1), pp.209-232.	No	Outcome is not binary		ESA Response
Avitabile, Ciro, and Rafael E. De Hoyos Navarro. 2015. The Heterogeneous Effect of Information on Student Performance: Evidence from a Randomized Control Trial in Mexico. *World Bank Working Paper*.	Yes			Initial Search
Banerjee, R., 2016. Corruption, norm violation and decay in social capital. *Journal of Public Economics*, 137, pp.14-27.	No	Outcome is not binary		ESA Response
Bao, J. and Ho, B., 2015. Heterogeneous effects of informational nudges on pro-social behavior. *The BE Journal of Economic Analysis & Policy*, 15(4), pp.1619-1655.	No	Not an empirical paper, only theory.		ESA Response
Beshears, J., Choi, J.J., Laibson, D., Madrian, B.C. and Milkman, K.L., 2015. The effect of providing peer information on retirement savings decisions. *The Journal of Finance*, 70(3), pp.1161-1201.	Yes			Initial Search
Bettinger, E.P., Long, B.T., Oreopoulos, P. and Sanbonmatsu, L., 2012. The role of application assistance and information in college decisions: Results from the H&R Block FAFSA experiment. *The Quarterly Journal of Economics*, *127*(3), pp.1205-1242.	Yes			Initial Search
Bhargava, S. and Manoli, D., 2015. Psychological frictions and the incomplete take-up of social benefits: Evidence from an IRS field experiment. *The American Economic Review*, *105*(11), pp.3489-3529.	Yes			Initial Search
Brown, M., Trautmann, S.T. and Vlahu, R., 2016. Understanding bank-run contagion. *Management Science*.	Yes			ESA Response
Cai, H., Chen, Y. and Fang, H., 2009. Observational learning: Evidence from a randomized natural field experiment. *The American Economic Review*, *99*(3), pp.864-882.	Yes			Initial Search

Caplin, A. and Martin, D.J., 2012. *Defaults and attention: The drop out effect* (No. w17988). National Bureau of Economic Research.	No	Not an information provision experiment	There is no zero information control group.	ESA Response
E., 2012. Inequality at work: The effect of peer salaries on job satisfaction. The American Economic Review, 102(6), pp.2981-3003.	No	Outcomes are hypothetical		Initial Search
Chen, Y., Harper, F.M., Konstan, J. and Xin Li, S., 2010. Social comparisons and contributions to online communities: A field experiment on movielens. *The American economic review*, *100*(4), pp.1358-1398.	No	No main outcome is binary.		ESA Response
Cialdini, R.B., Demaine, L.J., Sagarin, B.J., Barrett, D.W., Rhoads, K. and Winter, P.L., 2006. Managing social norms for persuasive impact. *Social influence*, *1*(1), pp.3-15.	No	Not an information provision experiment	There is no zero information control group.	Initial Search
Clark, R.L., Maki, J.A. and Morrill, M.S., 2014. Can Simple Informational Nudges Increase Employee Participation in a 401 (k) Plan?. *Southern Economic Journal*, *80*(3), pp.677-701.	Yes			Initial Search
Coffman, L.C., Featherstone, C.R. and Kessler, J.B., 2017. Can Social Information Affect What Job You Choose and Keep?. *American Economic Journal: Applied Economics*, *9*(1), pp.96-117.	Yes			Initial Search
Cooper, D.J. and Kagel, J.H., 2016. A failure to communicate: an experimental investigation of the effects of advice on strategic play. *European Economic Review*, *82*, pp.24-45.	No	Not an information provision experiment	Treatment is advice; might be interpreted differently by different subjects	ESA Response
d'Adda, G., Capraro, V. and Tavoni, M., 2017. Push, don't nudge: Behavioral spillovers and policy instruments. *Economics Letters*, *154*, pp.92-95.	No	Not an information provision experiment	Behavior in second round could be result of information or first round.	ESA Response
Damgaard, M.T. and Gravert, C., 2017. Now or never! The effect of deadlines on charitable giving: Evidence from two natural field experiments. *Journal of Behavioral and Experimental Economics*, *66*, pp.78-87.	No	Not information intervention		ESA Response
Damgaard, M.T. and Gravert, C., 2016. The hidden costs of nudging: Experimental evidence from reminders in fundraising. *Working paper*.	No	Not information intervention		ESA Response

Del Carpio, L., 2013. Are the Neighbors Cheating? Evidence from a Social Norm Experiment on Property Taxes in Peru. *Working paper*.	Yes			Initial Search
Dengler-Roscher, K., Estner, C. and Roscher, T., 2015. Nudging Academics to Didactic Training. *Working paper*.	Yes			ESA Response
Fellner, G., Sausgruber, R. and Traxler, C., 2013. Testing enforcement strategies in the field: Threat, moral appeal and social information. *Journal of the European Economic Association*, *11*(3), pp.634-660.	Yes			Initial Search
Frey, B.S. and Meier, S., 2004. Social comparisons and pro-social behavior: Testing" conditional cooperation" in a field experiment. *The American Economic Review*, *94*(5), pp.1717-1722.	Yes			Initial Search
Frydman, C. and Camerer, C., 2016. Neural evidence of regret and its implications for investor behavior. *The Review of Financial Studies*, *29*(11), pp.3108-3139.	No	Information provided is too informative.	Information is the value of a tradeable asset	ESA Response
Frydman, C. and Rangel, A., 2014. Debiasing the disposition effect by reducing the saliency of information about a stock's purchase price. *Journal of economic behavior & organization*, *107*, pp.541-552.	No	Not information intervention	Varies salience rather than information	ESA Response
Goldstein, N.J., Cialdini, R.B. and Griskevicius, V., 2008. A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of consumer Research*, *35*(3), pp.472-482.	Yes			Initial Search
Guillén, P. and Hakimov, R., 2015. *How to get truthful reporting in matching markets: A field experiment* (No. SP II 2015-208). WZB Discussion Paper.	No	Information provided is too informative.	Provide information on dominant strategy	ESA Response
Guillen, P. and Hing, A., 2014. Lying through their teeth: Third party advice and truth telling in a strategy proof mechanism. *European Economic Review*, *70*, pp.178-185.	No	Not information intervention		ESA Response
Hallsworth, M., List, J.A., Metcalfe, R.D. and Vlaev, I., 2017. The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics*, *148*, pp.14-31.	Yes			Initial Search

Hastings, J., Neilson, C.A. and Zimmerman, S.D., 2015. *The effects of earnings disclosure on college enrollment decisions* (No. w21300). National Bureau of Economic Research.	Yes			Initial Search
Hastings, J.S. and Weinstein, J.M., 2008. Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly journal of economics*, *123*(4), pp.1373-1414.	Yes			Initial Search
Ho, B., Taber, J., Poe, G. and Bento, A., 2016. The effects of moral licensing and moral cleansing in contingent valuation and laboratory experiments on the demand to reduce externalities. *Environmental and Resource Economics*, *64*(2), pp.317-340.	No	Outcome is not binary		ESA Response
Jensen, R., 2010. The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, *125*(2), pp.515-548.	Yes			Initial Search
Karadja, M., Mollerstrom, J. and Seim, D., 2017. Richer (and holier) than thou? The effect of relative income improvements on demand for redistribution. *Review of Economics and Statistics*.	No	Outcomes are hypothetical		Initial Search
Klinowski, D., 2015. Reluctant donors and their reactions to social information. *Working paper*.	No	Not information intervention	The utility of taking the action is not monotonic in the signal; it is U-shaped.	ESA Response
Krupka, E. and Weber, R.A., 2009. The focusing and informational effects of norms on pro-social behavior. *Journal of Economic Psychology*, *30*(3), pp.307-320.	Yes			Initial Search
Kuziemko, I., Norton, M.I., Saez, E. and Stantcheva, S., 2015. How elastic are preferences for redistribution? Evidence from randomized survey experiments. *The American Economic Review*, *105*(4), pp.1478-1508.	No	Outcomes are hypothetical		Initial Search
Lefgren, L.J., Sims, D.P. and Stoddard, O.B., 2016. Effort, luck, and voting for redistribution. *Journal of Public Economics*, *143*, pp.89-97.	No	Outcome is not binary		ESA Response
Liebman, J.B. and Luttmer, E.F., 2015. Would people behave differently if they better understood social security? Evidence from a field experiment. *American Economic Journal: Economic Policy*, 7(1), pp.275-299.	Yes			Initial Search
Lupia, A. and McCubbins, M.D., 1998. The democratic dilemma.	No	Information provided is too informative.		ESA Response

Moreno, Bernardo, Maria del Pino Ramos-Rosa, and Ismael Rodriguez-Lara. 2015. Conformity, information and truthful voting. *Working paper*.	No	Not information intervention	ESA Response
Nguyen, Trang. 2008. Information, Role Models, and Perceived Returns to Education: Experimental Evidence from Madagascar. *Working paper*	No	Outcome is not binary	Initial Search
Preece, J. and Stoddard, O., 2015. Why women don't run: Experimental evidence on gender differences in political competition aversion. *Journal of Economic Behavior & Organization*, *117*, pp.296-308.	No	Not information intervention	ESA Response
Servátka, M., 2009. Separating reputation, social influence, and identification effects in a dictator game. *European Economic Review*, *53*(2), pp.197-209.	No	Outcome is not binary	ESA Response
Tasoff, J. and Letzler, R., 2014. Everyone believes in redemption: Nudges and overoptimism in costly task completion. *Journal of Economic Behavior & Organization*, *107*, pp.107-122.	Yes		ESA Response
Wei, Shanshan, 2015. Social Influence in Charitable Giving. *Working paper*.	Yes		ESA Response
Zhu, Min, 2015. Experience Transmission: Truth-telling Adoption in Matching. *Working paper*.	No	Outcome is not binary	ESA Response

Notes: The **Included** column indicates whether the paper is one of the 22 papers that we include in our analysis. The **Source** column indicates whether the paper was one we found on our own "Initial Search" or one that was suggested to us in response to our email to the ESA listserve "ESA Response" (see details on the ESA listserve email below). There are five papers that we identified in our initial search that we ended up not including in our analysis. We initially considered including Card et al. (2012), Karadja et al. (2014), and Kuziemko et al. (2015) until we decided to limit our analysis to outcomes that were actions (e.g., rather than outcomes that were beliefs or responses to hypothetical questions). We had planned to include Nguyen (2008) until we recognized that the outcome variable was not binary. We had planned to include Cialdini et al. (2006) until we recognized that it did not have a no-information control group.

B Survey on Current Intuitions Appendix

Here is a copy of the survey we gave at the BSPA.

1. A policymaker friend wants to encourage older women to get mammograms. To this end, she plans to run a campaign to give accurate, quantitative information on the value of breast cancer screenings. Her research suggests most women who learn this information will increase their perceived value of a mammogram. She has enough money to run the information campaign in one of two sites, which are identical other than the number of women who got a mammogram last year.

If she simply wants to maximize the chance that her campaign increases the number of women getting a mammogram *this year*, what is her best approach? (select one)

☐ Do the information campaign at Site 1 where 10% of women got a mammogram *last year*.
☐ Do the information campaign at Site 2 where 75% of women got a mammogram *last year*.
☐ Do the information campaign at either site; it is equally likely to work at both sites.

2. (Optional) Why did you select the district you did? (free response)

_____

_____

_____

3. How would you describe yourself? (select all that apply)

☐ Policymaker
☐ Psychologist
☐ Economist
☐ Political Scientist
☐ Sociologist
☐ Other Academic: _____
☐ Other: _____

Figure B.1: Survey distributed at Behavioral Science and Policy Association Annual Conference 2018

C Theory Appendix

C.1 Individual-level model

In this section, we will show how utility-relevant parameters besides the one being signaled can be incorporated into the model from the main text.

C.1.a The general model

The agent has a net-utility function, u, that depends on the realization of a random scalar, X, and a random vector, Z. It captures the utility of taking up minus the utility of not taking up. **Untreated**, the agent's net utility is $\mathbb{E}\left[u(X,Z)\right]$, and she takes up if and only if that **prior** expected utility weakly exceeds zero. If the agent is exposed to the realization, v, of the random variable N—that is, **treated**—her net utility is updated to $\mathbb{E}\left[u(X,Z)\,\middle|\,N=v\right]$, and she takes up if and only if that **posterior** expected utility weakly exceeds zero.

Nothing in the previous paragraph specifically captures the idea that N is a noisy signal about X. The following convexity assumption captures part of this intuition.

> **Assumption C.1.** *The posterior expected utility lies between the prior expected utility and the expected utility of an agent who believes that X is exactly equal to the signal, v. That is, for some $\gamma \in [0,1]$,*
>
> $$\mathbb{E}\left[u(X,Z)\,\middle|\,N=v\right] \;=\; (1-\gamma)\,\mathbb{E}\left[u(X,Z)\right] + \gamma\,\mathbb{E}\left[u(X,Z)\,\middle|\,X=v\right].$$
>
> *holds.*

Note that γ need not be the same for every realization of the signal, v. In other words, we are not assuming the posterior expected utility is linear in $\mathbb{E}\left[u(X,Z)\,\middle|\,X=v\right]$. Our condition is closer to the *updating towards the signal* condition of Chambers and Healy (2012) than the *posterior linearity* of Diaconis and Ylvisaker (1979). Intuitively, treatment causes the agent's expected utility to update *towards* what it would be if the signal were simply a revelation of X, but not all the way there.

C.1.b Excising the nuisance parameters

The vector Z contains things that are utility-relevant but not signaled by N. In other words, Z can be thought of as a vector of **nuisance parameters**. We can greatly limit the role of Z in our model through the following conditional-independence assumption.

> **Assumption C.2.** *Conditional on X, the nuisance-parameter vector, Z, and the signal, N, are independent. That is, $Z \perp N \mid X$.*

Intuitively, this simply means that to the agent who knows X with certainty, the signal, N, contains no more utility-relevant information, which is sensible is N is just a noisy signal of X. This assumption allows us to derive the following effective net utility by marginalizing out Z:

$$\begin{aligned} v(x) &\equiv \mathbb{E}\left[u(X,Z)\,\middle|\,X=x\right], \\ &= \mathbb{E}\left[u(X,Z)\,\middle|\,X=x,\,N=\nu\right]. \end{aligned}$$

The first line is a definition; the equivalence on the second is a direct consequence of Assumption C.2. Now, we can write our posterior and prior expected utilities as

$$\begin{aligned} \mathbb{E}\left[u(X,Z)\,\middle|\,N=\nu\right] &= \mathbb{E}\left[v(X)\,\middle|\,N=\nu\right] \quad \text{and} \\ \mathbb{E}\left[u(X,Z)\right] &= \mathbb{E}\left[v(X)\right]. \end{aligned}$$

These expressions follow directly from the law of iterated expectations. From here, we can develop the theory in the same way as Section 3.1.

C.2 Experiment-level model

C.2.a The formal setup

An experiment is a population of agents being potentially exposed to a common information nudge, ν. As such, we consider the threshold, belief, and update strength of an agent that is uniformly and randomly selected from the population. We model this agent with the random variable (Θ, M, E).

Formally, we consider a random vector on the **outcome space** $\Omega^{\Theta ME} \equiv \Omega^{\Theta} \times \Omega^{M} \times \Omega^{E}$, where $\Omega^{E} = [0,1]$ and $\Omega^{\Theta} = \Omega^{M} = (a,b)$ for some (potentially infinite) constants $b > a$.[1a] When we write Ω with two superscripts, we mean the product of the corresponding outcome spaces; e.g., $\Omega^{ME} = \Omega^{M} \times \Omega^{E}$. The **event space** is $\mathscr{B}(\Omega^{\Theta ME})$, the Borel algebra associated with $\Omega^{\Theta ME}$. On the measurable space $\left(\Omega^{\Theta ME}, \mathscr{B}(\Omega^{\Theta ME})\right)$, the random vector (Θ, M, E)—whose components represent the threshold, prior, and update strength of a randomly drawn member of the population—is the identity function.

Given this setup, we characterize an **experiment** by its **nudge signal realization**, ν, and its **joint measure**, $F^{\Theta ME}$. The probability of (Θ, M, E) lying in any $C \in \mathscr{B}(\Omega^{\Theta ME})$ is given by the Lebesgue integral

$$\Pr\{(\Theta, M, E) \in C\} = \iiint_{C} dF^{\Theta ME}.$$

The **joint distribution function**, $F^{\Theta ME}(\theta, \mu, \varepsilon)$, equals the probability that (Θ, M, E) lies in the set $(a, \theta] \times (a, \mu] \times [0, \varepsilon]$. Note the slight abuse of notation: when the argument of F is a *set*, it should be treated as a *measure*; when its argument is a *point*, it should be treated as the corresponding *distribution function*.

1a. Setting Ω^{M} and Ω^{Θ} equal to the open set (a,b)—rather than the closed set $[a,b]$—prevents the potential complication of infinite realizations of M and Θ. When a or b is finite, it also prevents realizations of M and Θ from taking on those finite boundary values.

C.2.b Baseline, treatment effect, and the random vector (Θ, M, E)

The fraction of agents who take-up without being nudged—i.e., the **baseline**—is defined by

$$\beta \equiv \Pr\{\Theta \leq M\}, \tag{C.17}$$

while the fraction who take-up when nudged equals $\Pr\{\Theta \leq M + E(\nu - M)\}$. (The inequalities are weak because we assume the agent takes-up when her belief equals her threshold.) We define the **exact treatment effect**, τ^e, to be the change in the take-up rate caused by the nudge, that is

$$\tau^e \;\equiv\; \Pr\{M < \Theta \leq M + E(\nu - M)\} \;-\; \Pr\{M + E(\nu - M) < \Theta \leq M\}. \tag{C.18}$$

The two probabilities represent agents who are nudged *into* and *out of* take-up, respectively. Note that ν must *exceed* the realization of M for those nudged into take-up and must *be exceeded by* the realization of M for those nudged out of take-up.

Given this setup, the exact treatment effect can be written in terms of the conditional distribution function, $F^{\Theta|ME}$, as

$$\tau^e = \iint_{\Omega^{ME}} \left\{ F^{\Theta|ME}\left(\mu + \varepsilon(\nu - \mu) \,\middle|\, \mu, \varepsilon\right) - F^{\Theta|ME}\left(\mu \,\middle|\, \mu, \varepsilon\right) \right\} dF^{ME}(\mu, \varepsilon). \tag{C.19}$$

Note the integrand omits the probability that Θ is *exactly* equal to the smaller of μ and $\mu + \varepsilon(\nu - \mu)$, conditional on $(M, E) = (\mu, \varepsilon)$. This lines up with the strict and weak inequalities in Equation C.18. Figure C.2 shows those who contribute to the treatment effect (for a fixed E).

C.2.c The approximation

From here, we hope to approximate Equation C.19 when most update strengths are small. The following assumption will prove crucial to our approach.

Assumption C.3. *There is a strictly positive constant, $\bar{\delta} > 0$, such that for any realization, $(M, E) = (\mu, \varepsilon)$, the conditional distribution, $F^{\Theta|ME}\left(\theta \,\middle|\, \mu, \varepsilon\right)$, admits a density (with respect to Lebesgue measure), $f^{\Theta|ME}\left(\theta \,\middle|\, \mu, \varepsilon\right)$, on the interval $\Omega^{\Theta|M}(\mu)$ defined by*

$$\Omega^{\Theta|M}(\mu) \equiv \begin{cases} (a, \mu + \bar{\delta}] & \text{if } \mu < a + \bar{\delta}, \\ [\mu - \bar{\delta}, b) & \text{if } \mu > b - \bar{\delta}, \\ [\mu - \bar{\delta}, \mu + \bar{\delta}] & \text{otherwise.} \end{cases}$$

That is, for any $(\mu, \varepsilon) \in \Omega^{ME}$ and any $A \in \mathscr{B}\left(\Omega^{\Theta|M}(\mu)\right)$, it is true that

$$\int_A f^{\Theta|ME}\left(\theta \,\middle|\, \mu, \varepsilon\right) d\theta = \int_A dF^{\Theta|ME}\left(\theta \,\middle|\, \mu, \varepsilon\right). \tag{C.20}$$

Further, assume

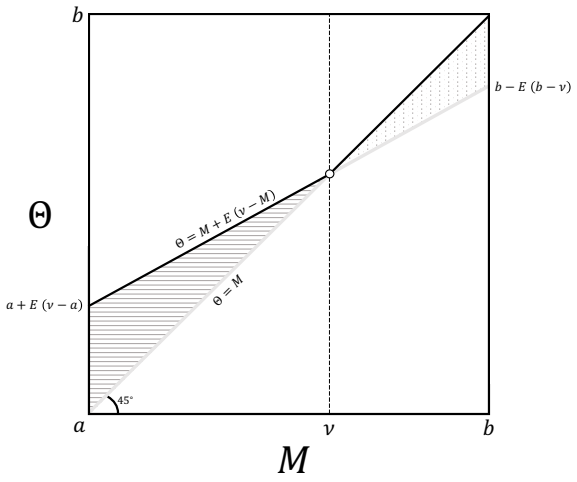


Figure C.2: Contributors to τ^e for a fixed E

NOTES: Agents in the hatched area are nudged into take up; agents in the dotted area are nudged out of take up. Dark boundaries are included in their adjacent regions; lighter ones are not. The hollow point at $M = v$ reflects agents who don't update and hence don't contribute to the treatment effect.

1. *the function* $f^{\Theta|ME}(\theta \,|\, \mu, \varepsilon)$ *is equal to zero for any* $\theta \notin \Omega^{\Theta|M}(\mu)$*, and*
2. *the family of conditional densities,* $\left\{f^{\Theta|ME}(\mu+\delta \,|\, \mu, \varepsilon)\right\}_{(\mu,\varepsilon)\in\Omega^{ME}}$*, is equicontinuous at* $\delta = 0$*.*

Figure C.3 illustrates the region within $\Omega^{\Theta M}$ where $f^{\Theta|ME}$ acts as a density for the measure $F^{\Theta|ME}$. The first of the final two conditions in Assumption C.3 sets $f^{\Theta|ME}$ equal to zero outside of the range in the illustration, which serves merely to simplify future formulas. The second condition selects a particular, well-behaved *version* of the density which allows us to approximate the integrand in Equation C.19 with $\varepsilon(v-\mu) f^{\Theta|ME}(\mu \,|\, \mu, \varepsilon)$ when updates are small.[2a] This leads us to define the **approximate treatment effect** as

$$\tau \equiv \iint_{\Omega^{ME}} \varepsilon(v-\mu) f^{\Theta|ME}(\mu \,|\, \mu, \varepsilon)\, dF^{ME}(\mu, \varepsilon). \tag{C.21}$$

For now, we simply assume this definition is meaningful; later we will provide more intuition.

2a. Technically, we are ruling out a version of the density where we arbitrarily set the value of $f^{\Theta|ME}(\theta \,|\, \mu, \varepsilon)$ for all (μ, ε) when $\theta = \mu$.

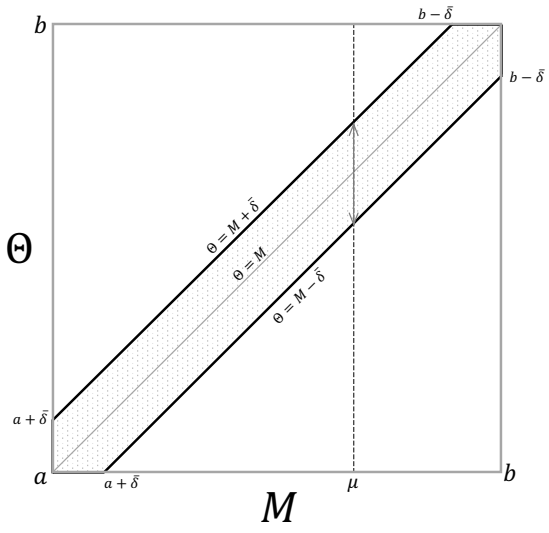


Figure C.3: Realizations and conditional densities for Θ

NOTES: The dotted region shows where the density $f^{\Theta|ME}$ represents the measure $F^{\Theta|ME}$. Off of this area, $f^{\Theta|ME}$ is a zero-valued function.

Assumption C.4. *The integrand in Equation C.21 is F^{ME}-integrable. That is, the integral*

$$\iint_{\Omega^{ME}} \varepsilon\,|\nu-\mu|\,f^{\Theta|ME}\left(\mu\,\middle|\,\mu,\varepsilon\right)\,dF^{ME}(\mu,\varepsilon)$$

exists and is finite.

Going forward, we will show that the approximation in Equation C.21 is, in fact, a good one by bounding $|\tau^e-\tau|$. To do so, we will rely on a simple observation: each agent represented by the integrand in Equation C.21 has equal prior and threshold. We call such agents **marginal**; understanding them is key to what follows.

C.2.d Marginal agents

To formalize the idea of marginality, we define an auxiliary random variable, the **deficit**, as $\Delta \equiv \Theta - M$. Marginal agents have deficits equal to zero. Across the population, the deficit's outcome can be anywhere on the interval $(a-b,\, b-a)$, which we will call Ω^{Δ}.[3a] Conditional on $(M,E)=(\mu,\varepsilon)$, however, the deficit's outcome is restricted to $\Delta \in \Omega^{\Theta}-\mu$. Hence, the realization of

3a. The same caveats from Footnote 1a also hold with the boundaries of Ω^{Δ}.

(Δ, M) must be on the set

$$\{(\delta, \mu) \in \Omega^{\Delta M} \ : \ \delta \in \Omega^{\Theta} - \mu\},$$

which is illustrated in Figure C.4. Note that this is a strict subset of $\Omega^{\Delta M}$.

Going forward, we will often characterize an experiment by the joint measure $F^{\Delta ME}$ instead of $F^{\Theta ME}$. So long as the measure $F^{\Delta ME}$ only places weight on the set illustrated in Figure C.4, these two approaches are isomorphic since, for any sets $A \in \mathscr{B}(\Omega^{\Delta ME})$ and $B \in \mathscr{B}(\Omega^{\Theta ME})$, we can write

$$\begin{aligned} F^{\Delta ME}(A) &= F^{\Theta ME}\left(\{(\delta + \mu, \mu, \varepsilon) \cap \Omega^{\Theta ME} : (\delta, \mu, \varepsilon) \in A\}\right), \text{ and} \\ F^{\Theta ME}(B) &= F^{\Delta ME}\left(\{(\theta - \mu, \mu, \varepsilon) \quad\quad\quad\quad : (\theta, \mu, \varepsilon) \in B\}\right). \end{aligned}$$

(The intersection in the first equation serves to ensure that points in $\Omega^{\Delta M}$, but outside the set illustrated in Figure C.4, are excluded.)

Un-nudged, an agent takes up if and only if $\Delta \leq 0$; nudged, she takes up if and only if $\Delta \leq E(v - M)$. We can hence rewrite Equations C.17, C.18, and C.19 as

$$\beta = \Pr\{\Delta \leq 0\}, \tag{C.17′}$$

$$\tau^e = \Pr\{0 < \Delta \leq E(v - M)\} - \Pr\{E(v - M) < \Delta \leq 0\}, \text{ and} \tag{C.18′}$$

$$\tau^e = \iint_{\Omega^{ME}} \left\{F^{\Delta|ME}\left(\varepsilon(v - \mu) \,\middle|\, \mu, \varepsilon\right) - F^{\Delta|ME}\left(0 \,\middle|\, \mu, \varepsilon\right)\right\} dF^{ME}(\mu, \varepsilon). \tag{C.19′}$$

We can also rewrite Assumption C.3 in terms of such agents.

Result C.1. *For any $(\mu, \varepsilon) \in \Omega^{ME}$, define*

$$f^{\Delta|ME}\left(\delta \,\middle|\, \mu, \varepsilon\right) \equiv f^{\Theta|ME}\left(\mu + \delta \,\middle|\, \mu, \varepsilon\right),$$

and

$$\Omega^{\Delta|M}(\mu) \equiv \begin{cases} (a - \mu, \bar{\delta}] & \text{if } \mu < a + \bar{\delta}, \\ [-\bar{\delta}, b - \mu) & \text{if } \mu > b - \bar{\delta}, \\ [-\bar{\delta}, \bar{\delta}] & \text{otherwise.} \end{cases}$$

Then, the function $f^{\Delta|ME}\left(\delta \,\middle|\, \mu, \varepsilon\right)$ is a density for the conditional measure, $F^{\Delta|ME}\left(\cdot \,\middle|\, \mu, \varepsilon\right)$. That is, for any $(\mu, \varepsilon) \in \Omega^{ME}$ and $A \in \mathscr{B}(\Omega^{\Delta|M}(\mu))$, it is true that

$$\int_A f^{\Delta|ME}\left(\delta \,\middle|\, \mu, \varepsilon\right) d\delta = \int_A dF^{\Delta|ME}\left(\delta \,\middle|\, \mu, \varepsilon\right).$$

Further,

1. *the function $f^{\Delta|ME}\left(\delta \,\middle|\, \mu, \varepsilon\right)$ is equal to zero for any $\delta \notin \Omega^{\Delta|M}(\mu)$, and*
2. *the family of conditional densities, $\left\{f^{\Delta|ME}\left(\delta \,\middle|\, \mu, \varepsilon\right)\right\}_{(\mu,\varepsilon)\in\Omega^{ME}}$, is equicontinuous at $\delta = 0$.*

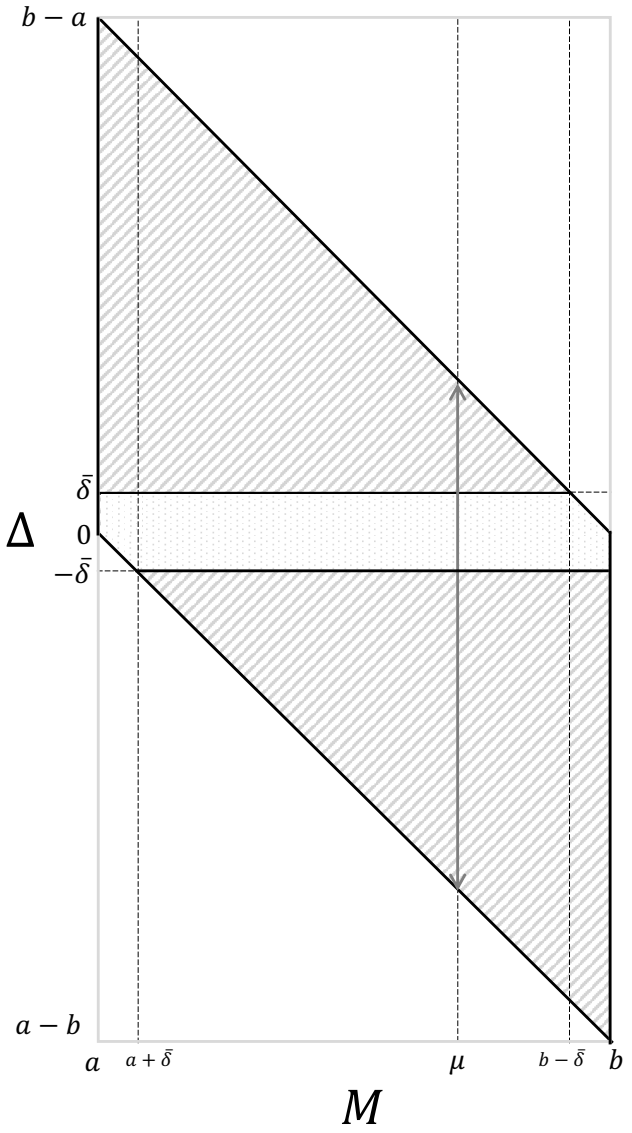


Figure C.4: Realizations and conditional densities for Δ

NOTES: The union of the dotted and the hatched regions shows where (Δ, M) realizations can occur; the dotted region shows where the density $f^{\Delta|ME}$ represents the measure $F^{\Delta|ME}$. Off of this area, $f^{\Delta|ME}$ is a zero-valued function.

Proof. For the main result, note that, conditional on $(M,E) = (\mu,\varepsilon)$, the probability that $\Delta \in A$ is equal to the probability that $\Theta \in A+\mu$. That is,

$$\int_A dF^{\Delta|ME}(\delta \,|\, \mu,\varepsilon) = \int_{A+\mu} f^{\Theta|ME}(\theta \,|\, \mu,\varepsilon)\, d\theta.$$

A simple change of variables shows that

$$\int_{A+\mu} f^{\Theta|ME}(\theta \,|\, \mu,\varepsilon)\, d\theta = \int_A f^{\Theta|ME}(\delta+\mu \,|\, \mu,\varepsilon)\, d\delta.$$

Stringing together the two equations establishes the result. The final two results follow immediately by definition. □

Figure C.4 illustrates the region within $\Omega^{\Delta M}$ where $f^{\Delta|ME}$ is a density for the measure $F^{\Delta|ME}$. Given Result C.1, we can also rewrite Equation C.21 as

$$\tau = \iint_{\Omega^{ME}} \varepsilon\,(\nu-\mu)\, f^{\Delta|ME}(0 \,|\, \mu,\varepsilon)\, dF^{ME}(\mu,\varepsilon) \qquad \text{(C.21}'\text{)}$$

and Assumption C.4 as a statement about the integrability of the integrand in Equation C.21′.

So, Result C.1 gives a convenient expression for the measure $F^{\Delta|ME}$. It also gives convenient expressions for the measures F^{Δ} and $F^{ME|\Delta}$.

Result C.2. *Define the function $f^{\Delta}(\delta)$ by*

$$f^{\Delta}(\delta) \equiv \iint_{\Omega^{ME}} f^{\Delta|ME}(\delta \,|\, \mu,\varepsilon)\; dF^{ME}(\mu,\varepsilon),$$

Then f^{Δ} is a density for the marginal measure, F^{Δ}, on $\mathscr{B}([-\bar{\delta},\bar{\delta}])$. That is, for any $A \in \mathscr{B}([-\bar{\delta},\bar{\delta}])$, it is true that

$$\int_A f^{\Delta}(\delta)\, d\delta = \int_A dF^{\Delta}(\delta).$$

Further, $f^{\Delta}(\delta)$ is continuous at $\delta = 0$.

Proof. Using the definition above, we have

$$\begin{aligned}\int_A f^{\Delta}(\delta)\, d\delta &= \int_A \left\{ \iint_{\Omega^{ME}} f^{\Delta|ME}(\delta \,|\, \mu,\varepsilon)\, dF^{ME}(\mu,\varepsilon) \right\} d\delta, \\ &= \iint_{\Omega^{ME}} \left\{ \int_A f^{\Delta|ME}(\delta \,|\, \mu,\varepsilon)\, d\delta \right\} dF^{ME}(\mu,\varepsilon),\end{aligned}$$

where the first equation comes from the definition of f^{Δ} in the statement of the result, and the second comes from a Fubini–Tonelli change of integration order. We can continue with

$$\begin{aligned}\int_A f^{\Delta}(\delta)\, d\delta &= \iint_{\Omega^{ME}} \left\{ \int_A dF^{\Delta|ME}(\delta \,|\, \mu,\varepsilon) \right\} dF^{ME}(\mu,\varepsilon), \\ &= \iiint_{A\times\Omega^{ME}} dF^{\Delta ME}(\delta,\mu,\varepsilon), \\ &= \iiint_{\Delta^{-1}(A)} dF^{\Delta ME}(\delta,\mu,\varepsilon), \\ &= \int_A dF^{\Delta}(\delta).\end{aligned}$$

The first equation follows from Result C.1, while the second follows from the definition of a conditional distribution. The third follows because the integration region in the second integral is just $\Delta^{-1}(A)$, and the fourth follows from the definition of a marginal distribution. So we have shown that f^{Δ}, as defined above, is indeed a density for the marginal distribution of Δ.

To see that $f^{\Delta}(\delta)$ is continuous at $\delta = 0$, we use the equicontinuity part of Result C.1, which says that for any $x > 0$, there is a $y > 0$ such that when $|\delta| < y$, $\left| f^{\Delta|ME}\left(\delta \middle| \mu, \varepsilon\right) - f^{\Delta|ME}\left(0 \middle| \mu, \varepsilon\right) \right| < x$ for all $(\mu, \varepsilon) \in \Omega^{ME}$. This immediately implies that when $|\delta| < y$, $\left| f^{\Delta}(\delta) - f^{\Delta}(0) \right| < x$, which is the definition of continuity at zero for f^{Δ}. □

The continuity result shows that any version of the density f^{Δ} that meets the definition above has the same value at zero.

Going forward, we use the versions of $f^{\Delta|ME}$ and f^{Δ} listed in Results C.1 and C.2.

Proposition C.1. *Define the measure $F^{ME|\Delta}$ such that, for any $\delta \in [-\bar{\delta}, \bar{\delta}]$ and $B \in \mathscr{B}\left(\Omega^{ME}\right)$,*

$$\iint_B dF^{ME|\Delta}\left(\mu, \varepsilon \middle| \delta\right) \equiv \begin{cases} \dfrac{1}{f^{\Delta}(\delta)} \iint_B f^{\Delta|ME}\left(\delta \middle| \mu, \varepsilon\right) dF^{ME}(\mu, \varepsilon) & \text{if} \quad f^{\Delta}(\delta) > 0, \\ \iint_B dF^{ME}(\mu, \varepsilon) & \text{if} \quad f^{\Delta}(\delta) = 0. \end{cases}$$

Then, for realizations of Δ on $[-\bar{\delta}, \bar{\delta}]$, $F^{ME|\Delta}$ is a version of the measure of (M, E) conditional on Δ. That is, for any $A \in \mathscr{B}\left([-\bar{\delta}, \bar{\delta}]\right)$ and $B \in \mathscr{B}\left(\Omega^{ME}\right)$, it is true that

$$\int_A \left\{ \iint_B dF^{ME|\Delta}\left(\mu, \varepsilon \middle| \delta\right) \right\} f^{\Delta}(\delta)\, d\delta = \iiint_{A \times B} dF^{\Delta ME}(\delta, \mu, \varepsilon). \tag{C.22}$$

Further, for any $B \in \mathscr{B}\left(\Omega^{ME}\right)$, if $f^{\Delta}(0) > 0$, then the function defined by the integral

$$\iint_B dF^{ME|\Delta}\left(\mu, \varepsilon \middle| \delta\right)$$

is continuous at $\delta = 0$.

Proof. Define $A_+ \equiv \{\delta \in A : f^{\Delta}(\delta) > 0\}$. On $A \setminus A_+$, the left-hand side of the Equation C.22 equals zero, regardless of how $F^{ME|\Delta}$ is defined. The right-side does as well, since

$$0 \leq \iiint_{(A \setminus A_+) \times B} dF^{\Delta ME}(\delta, \mu, \varepsilon) \leq \iiint_{(A \setminus A_+) \times \Omega^{ME}} dF^{\Delta ME}(\delta, \mu, \varepsilon) = \int_{A \setminus A_+} dF^{\Delta}(\delta) = 0.$$

So, we only need show that the equation holds on A_+. Starting with the left-hand side of Equation C.22, we can proceed as follows:

$$\begin{aligned}
&\int_{A_+}\left\{\iint_B f^{\Delta|ME}(\delta\,|\,\mu,\varepsilon)\,dF^{ME}(\mu,\varepsilon)\right\}d\delta \\
&\qquad = \iint_B\left\{\int_{A_+} f^{\Delta|ME}(\delta\,|\,\mu,\varepsilon)\,d\delta\right\}dF^{ME}(\mu,\varepsilon), \\
&\qquad = \iint_B\left\{\int_{A_+} dF^{\Delta|ME}(\delta\,|\,\mu,\varepsilon)\right\}dF^{ME}(\mu,\varepsilon), \\
&\qquad = \iiint_{A_+\times B} dF^{\Delta ME}(\delta,\mu,\varepsilon).
\end{aligned}$$

where the first expression comes from plugging in our version of $F^{ME|\Delta}$ and cancelling the $f^{\Delta}(\delta)$ terms, the first equality comes from a Fubini–Tonelli interchange of integration order, the second comes from the fact that $f^{\Delta|ME}$ is a density for $F^{\Delta|ME}$, and the third follows from the definition of conditional distribution applied to $F^{\Delta|ME}$ and F^{ME}. So, we see that our equation does indeed satisfy the requirements to be a version of the conditional measure $F^{ME|\Delta}$.

The proof of the continuity result at the end is almost identical to the proof of the continuity result from Result C.2. □

Note that the proof shows that the value taken on by $F^{ME|\Delta}(\mu,\varepsilon\,|\,\delta)$ for values of δ where $f^{\Delta}(\delta)=0$ doesn't matter; our choice, $F^{ME}(\mu,\varepsilon)$, is an arbitrarily chosen probability measure on Ω^{ME}. The continuity result shows that any version of the density f^{Δ} that meets the definition will yield equal expectations conditional on $\Delta=0$.

Going forward, for realizations of Δ on $[-\bar{\delta},\bar{\delta}]$, we use the version of $F^{ME|\Delta}$ listed in Proposition C.1.

To summarize, we assume densities for $F^{\Theta|ME}$ that let us construct intuitive expressions for the measures $F^{\Delta|ME}$, $F^{ME|\Delta}$, and F^{Δ}. We leave the measure F^{ME} general, however, so it is free to include atoms, among other things. This allows us, for instance, to place an atom at $E=0$ to model a finite fraction of the population that ignores the nudge altogether.

C.2.e What drives τ?

Looking at Equation C.21′, the approximate treatment effect, τ, can be written as an unconditional expectation, that is,

$$\tau = \mathbb{E}\left[E\,(v-M)\,f^{\Delta|ME}(0\,|\,M,E)\right].$$

The importance of marginal agents is captured by the density term. Of course, it would be more intuitive for the entire expectation to be conditional on the marginal sub-population. Assuming $f^{\Delta}(0)$ is strictly positive, Proposition C.1 allows us to write

$$\tau = f^{\Delta}(0)\;\mathbb{E}\left[E\,(v-M)\,\middle|\,\Delta=0\right]. \tag{C.23}$$

In this light, Assumption C.4 becomes a statement about the existence of the moment $\mathbb{E}\left[E\,(v-M)\,\middle|\,\Delta=0\right]$ (i.e., the expected update among the marginal).[4a] Going forward, we will assume there are marginal agents. That is,

Assumption C.5. *The deficit density at zero—$f^{\Delta}(0)$—is finite and strictly positive.*

Intuitively then, τ is driven by three forces. First, it is driven by how many marginal agents there are, captured by the density $f^{\Delta}(0)$. Second, it is driven by how strongly the marginal agents update, captured by E in the conditional expectation. And third, it is driven by the discord between the information and the prior for marginal agents (the "newsworthiness" of the signal), captured by $v-M$ in the conditional expectation.

With a bit of algebra, we can also write τ as

$$\tau \;=\; f^{\Delta}(0)\;\mathbb{E}\left[E\,\middle|\,\Delta=0\right]\left\{v \;-\; \frac{\mathbb{E}\left[E\,M\,\middle|\,\Delta=0\right]}{\mathbb{E}\left[E\,\middle|\,\Delta=0\right]}\right\}. \tag{C.24}$$

Tacitly, we are assuming that $\mathbb{E}\left[E\,\middle|\,\Delta=0\right]$ is strictly positive.

Assumption C.6. *The expected update strength of the marginal, $\mathbb{E}\left[E\,\middle|\,\Delta=0\right]$, is strictly positive.*

In Equation C.24 then, the strength of marginal updates is captured by $\mathbb{E}\left[E\,\middle|\,\Delta=0\right]$, while the "newsworthiness" of the signal comes from comparing it to $\mathbb{E}\left[E\,M\,\middle|\,\Delta=0\right]/\mathbb{E}\left[E\,\middle|\,\Delta=0\right]$, which is the average of marginal priors, weighted by update strength.

Note that when E and M aren't strongly correlated, the simple average prior among the marginal, $\mathbb{E}\left[M\,\middle|\,\Delta=0\right]$, becomes a good approximation for $\mathbb{E}\left[E\,M\,\middle|\,\Delta=0\right]/\mathbb{E}\left[E\,\middle|\,\Delta=0\right]$.

C.2.f Showing τ is a good approximation

We now return to showing that τ is a good approximation for τ^{e} by bounding the absolute approximation error, $|\tau^{e}-\tau|$. For any $X\in\mathscr{B}(\Omega^{ME})$, define

$$\tau^{e}(X) \equiv \iint_{X}\left\{F^{\Delta|ME}\left(\varepsilon\,(v-\mu)\,\middle|\,\mu,\varepsilon\right)-F^{\Delta|ME}\left(0\,\middle|\,\mu,\varepsilon\right)\right\}dF^{ME}(\mu,\varepsilon),$$
$$\tau(X) \equiv \iint_{X}\varepsilon\,(v-\mu)\;f^{\Delta|ME}\left(0\,\middle|\,\mu,\varepsilon\right)\,dF^{ME}(\mu,\varepsilon).$$

Our general strategy will be to partition Ω^{ME} into regions that can be handled by differing methods. To do this, we will need to strengthen Assumption C.4.

4a. The moment $\mathbb{E}\left[E\,(v-M)\,\middle|\,\Delta=0\right]$ is said to **exist** if the integral $\mathbb{E}\left[E\,|v-M|\,\middle|\,\Delta=0\right]$ exists and is finite.

Assumption C.7. *The moments* $\mathbb{E}\left[E^2(\nu-M)^2\right]$ *and* $\mathbb{E}\left[E^2(\nu-M)^2\,\middle|\,\Delta=0\right]$ *exist.*

Notice that, Assumptions C.5 and C.7 combine to imply Assumption C.4.[5a]

C.2.f.i Where Updates are Large

When updates are sufficiently large, the deficit density is not guaranteed to exist. This happens on the region given by

$$A \equiv \left\{(\mu,\varepsilon)\in\Omega^{ME} \;:\; \varepsilon\,|\nu-\mu|>\bar{\delta}\right\}. \tag{C.25}$$

The the error contributed on A can be easily bounded.

Result C.3. *The absolute error contributed on* A *obeys the following bound:*

$$\left|\tau^e(A)-\tau(A)\right| \;\leq\; \frac{\mathbb{E}\left[E^2(\nu-M)^2\right]}{\bar{\delta}^2} \;+\; \frac{f^{\Delta}(0)}{\bar{\delta}}\,\mathbb{E}\left[E^2(\nu-M)^2\,\middle|\,\Delta=0\right].$$

Proof. Our strategy is first to bound $|\tau^e(A)|$ and $|\tau(A)|$ separately and then to combine the bounds with the triangle inequality. Starting with $|\tau^e(A)|$, it is clear the magnitude of its integrand is bounded by 1. Hence,

$$\begin{aligned}
\left|\tau^e(A)\right| &\leq \iint_A dF^{ME}(\mu,\varepsilon),\\
&= \Pr\{(M,E)\in A\},\\
&= \Pr\left\{E\,|\nu-M|>\bar{\delta}\right\},\\
&\leq \frac{\mathbb{E}\left[E^2(\nu-M)^2\right]}{\bar{\delta}^2}.
\end{aligned}$$

The final inequality comes from Markov's inequality and the fact that $\Pr\left\{E\,|\nu-M|>\bar{\delta}\right\}=\Pr\left\{E^2(\nu-M)^2>\bar{\delta}^2\right\}$. Moving on to $|\tau(A)|$, we have

$$\begin{aligned}
|\tau(A)| &= f^{\Delta}(0)\left|\iint_A \varepsilon(\nu-\mu)\;dF^{ME|\Delta}\left(\mu,\varepsilon\,|\,0\right)\right|,\\
&= f^{\Delta}(0)\left|\iint_{\Omega^{ME}} \mathbb{1}_{\{(\mu,\varepsilon)\in A\}}\;\varepsilon(\nu-\mu)\;dF^{ME|\Delta}\left(\mu,\varepsilon\,|\,0\right)\right|,\\
&\leq f^{\Delta}(0)\sqrt{\iint_{\Omega^{ME}} \mathbb{1}^2_{\{(\mu,\varepsilon)\in A\}}\;dF^{ME|\Delta}\left(\mu,\varepsilon\,|\,0\right)}\;\sqrt{\iint_{\Omega^{ME}} \varepsilon^2(\nu-\mu)^2\;dF^{ME|\Delta}\left(\mu,\varepsilon\,|\,0\right)},\\
&= f^{\Delta}(0)\sqrt{\Pr\left\{(M,E)\in A\,\middle|\,\Delta=0\right\}}\;\sqrt{\mathbb{E}\left[E^2(\nu-M)^2\,\middle|\,\Delta=0\right]},\\
&= f^{\Delta}(0)\sqrt{\Pr\left\{E\,|\nu-M|>\bar{\delta}\,\middle|\,\Delta=0\right\}}\;\sqrt{\mathbb{E}\left[E^2(\nu-M)^2\,\middle|\,\Delta=0\right]},\\
&\leq f^{\Delta}(0)\sqrt{\frac{\mathbb{E}\left[E^2(\nu-M)^2\,\middle|\,\Delta=0\right]}{\bar{\delta}^2}}\;\sqrt{\mathbb{E}\left[E^2(\nu-M)^2\,\middle|\,\Delta=0\right]},\\
&= \frac{f^{\Delta}(0)}{\bar{\delta}}\,\mathbb{E}\left[E^2(\nu-M)^2\,\middle|\,\Delta=0\right].
\end{aligned}$$

5a. Given Assumption C.5, Assumption C.4 becomes equivalent to $\mathbb{E}\left[E\,|\nu-M|\,\middle|\,\Delta=0\right]<\infty$. To see this, start with the inequality $E^2(\nu-M)^2+1>E\,|\nu-M|$. (When $E\,|\nu-M|\leq 1$, the one on the right-hand side ensures the inequality holds. When $E\,|\nu-M|>1$, the $E^2\,|\nu-M|^2$ on the right-hand side ensures the inequality holds.) If we integrate it, we get $\mathbb{E}\left[E^2(\nu-M)^2\,\middle|\,\Delta=0\right]+1>\mathbb{E}\left[E\,|\nu-M|\,\middle|\,\Delta=0\right]$. Hence, if the left-hand side is finite, the right-hand side must be also.

The first inequality is the Cauchy–Schwarz inequality, while the second is the same Markov's inequality trick from before. From here, the triangle inequality establishes the desired bound. □

C.2.f.ii Where Updates are Smaller

Off the set A, Result C.1 states that the conditional deficit density exists. This means that, for any measurable $X \subseteq \Omega^{ME} \setminus A$, we can write

$$\begin{aligned}&\tau^e(X) - \tau(X)\\ &= \iint_X \left\{ \int_0^{\varepsilon(\nu-\mu)} \left\{ f^{\Delta|ME}(\delta \,|\, \mu,\varepsilon) - f^{\Delta|ME}(0 \,|\, \mu,\varepsilon) \right\} d\delta \right\} dF^{ME}(\mu,\varepsilon). \quad \text{(C.26)}\end{aligned}$$

Now, we bound $|\tau^e(X) - \tau(X)|$ for three important classes of X. As we go along, we will motivate these these classes with examples.

Example C.1 (Normal, Independent Thresholds)**.** *Let $\Omega^\Theta = (-\infty, \infty)$, and assume that the threshold, Θ, is distributed normally with mean and variance $(\mu_\Theta, \sigma_\Theta^2)$ and that it is independent of (M, E). Then, conditional on $(M, E) = (\mu, \varepsilon)$, the distribution of the deficit, Δ, is just the distribution of Θ with its mean shifted to the left by μ. That is, letting φ represent the standard normal density, we have*

$$f^{\Delta|ME}(\delta \,|\, \mu,\varepsilon) \;=\; \frac{1}{\sigma_\Theta}\, \varphi\left(\frac{\delta - (\mu_\Theta - \mu)}{\sigma_\Theta} \right).$$

This function is continuous in δ; in fact, it is Lipschitz continuous. To see this, note that

$$\partial_1 f^{\Delta|ME}(\delta \,|\, \mu,\varepsilon) = \frac{1}{\sigma_\Theta^2}\, \varphi'\left(\frac{\delta - (\mu_\Theta - \mu)}{\sigma_\Theta} \right).$$

Since $\varphi'(x) = -x\,\varphi(x)$ and $\varphi''(x) = (x^2 - 1)\,\varphi(x)$, we see that $\varphi'(x)$ attains its maximum magnitude when $x = \pm 1$. Hence, we have that $|\varphi'(x)| \leq |\varphi'(1)| = 1/\sqrt{2\pi e}$. The Mean Value Theorem then tells us that, for any δ_1 and δ_2, and any (μ, ε), we have

$$\left| f^{\Delta|ME}(\delta_2 \,|\, \mu,\varepsilon) - f^{\Delta|ME}(\delta_1 \,|\, \mu,\varepsilon) \right| \;\leq\; \frac{1}{\sigma_\Theta^2 \sqrt{2\pi e}}\; |\delta_2 - \delta_1|.$$

In fact, since the Lipschitz constant is independent of (μ, ε), we have that the family of functions $\{f^{\Delta|ME}(\delta \,|\, \mu,\varepsilon)\}_{(\mu,\varepsilon) \in \Omega^{ME}}$ is equi-Lipschitz continuous.

Looking to this example, our first class will be densities that obey an equi-Lipschitz condition.

Result C.4 (Equi-Lipschitz-Continuous Densities). *Fix a set $X \subseteq \Omega^{ME} \setminus A$, and for some $k > 0$, assume the inequality*

$$\left| f^{\Delta|ME}\left(\delta \,\middle|\, \mu, \varepsilon\right) - f^{\Delta|ME}\left(0 \,\middle|\, \mu, \varepsilon\right) \right| \;\leq\; k\,|\delta|$$

holds for any $(\mu, \varepsilon) \in X$ and any δ between 0 and $\varepsilon\,(\nu - \mu)$. Then, the following bound holds:

$$\left|\tau^e(X) - \tau(X)\right| \;\leq\; \frac{k}{2}\,\mathbb{E}\left[E^2\,(\nu - M)^2\right].$$

Proof. Using the assumed inequality, the desired bound is almost immediate:

$$\begin{aligned}
\left|\tau^e(X) - \tau(X)\right| &\leq k \iint_X \left| \int_0^{\varepsilon\,(\nu-\mu)} |\delta|, d\delta \right| dF^{ME}(\mu, \varepsilon), \\
&= \frac{k}{2} \iint_X \varepsilon^2\,(\nu - \mu)^2 \; dF^{ME}(\mu, \varepsilon), \\
&\leq \frac{k}{2} \iint_{\Omega^{ME}} \varepsilon^2\,(\nu - \mu)^2 \; dF^{ME}(\mu, \varepsilon), \\
&= \frac{k}{2}\,\mathbb{E}\left[E^2\,(\nu - M)^2\right].
\end{aligned}$$

The first inequality is assumed; the second follows because the integrand is non-negative. □

While the condition from the previous result is sensible, it can be violated. The density from Example C.1 worked mainly because its derivative was everywhere bounded. This can fail in two ways, as the next example shows.

Example C.2 (Beta-Distributed, Independent Thresholds). *Let $\Omega^\Theta = [0, 1]$, and assume that the threshold, Θ, is beta-distributed with strictly positive, real parameters (p, q) and that it is independent of (M, E). Then, conditional on $(M, E) = (\mu, \varepsilon)$, the deficit, Δ, has a shifted beta distribution. Letting* B *represent the beta function, we have*

$$f^{\Delta|ME}\left(\delta \,\middle|\, \mu, \varepsilon\right) \;=\; \frac{\left(\delta + \mu\right)^{p-1}\left(1 - \left(\delta + \mu\right)\right)^{q-1}}{\mathrm{B}(p, q)},$$

where the domain of $f^{\Delta|ME}$ is limited to $\delta \in [-\mu, 1 - \mu]$. Looking back to Example C.1, we see that our equi-Lipschitz condition fails if the first derivative is unbounded. This derivative is given by

$$\begin{aligned}
\partial_1 f^{\Delta|ME}\left(\delta \,\middle|\, \mu, \varepsilon\right) \;=\; \frac{1}{\mathrm{B}(p, q)} \Big\{ &(p-1)\left(\delta + \mu\right)^{p-2}\left(1 - \left(\delta + \mu\right)\right)^{q-1} \\
&- (q-1)\left(\delta + \mu\right)^{p-1}\left(1 - \left(\delta + \mu\right)\right)^{q-2} \Big\}.
\end{aligned}$$

As $\delta + \mu \to 0$, the derivative becomes unbounded if $p \in (0, 2) \setminus \{1\}$. For $p \in (0, 1)$, the density is unbounded as well, while for $p \in (1, 2)$, the density goes to zero. Similarly, as $\delta + \mu \to 1$, the derivative becomes unbounded if $q \in (0, 1) \setminus \{1\}$.

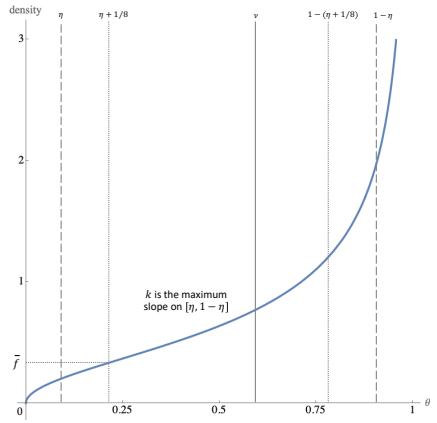


Figure C.5: Example C.2 threshold density for $(p, q) = (3/2, 1/2)$.

NOTES: The derivative becomes unbounded when θ approaches either 0 or 1, but the density only becomes unbounded when θ approaches 1. These two cases correspond to Results C.5 and C.6. (The labeled horizontal and vertical lines correspond to points of interest from the standard partition constructed near the end of Section C.2.g.)

For $q \in (0, 1)$, the density is unbounded as well, while for $q \in (1, 2)$, the density goes to zero.

So, the derivative can be unbounded in two ways: when the density is unbounded and when it is not. The threshold density with parameters $(p, q) = (3/2, 1/2)$, plotted in Figure C.5, illustrates this distinction. At zero, the density is zero, but the slope is infinite. At one, the slope and density are both infinite (i.e., there is a vertical asymptote).

The distinction in the previous example informs the other two classes of sets on which we will bound the absolute error.

C.2.f.iii Bounded threshold density.

Whenever the densities on X are bounded, the absolute error is easy to bound.

Result C.5. *Fix a set $X \subseteq \Omega^{ME} \setminus A$, and for some $\bar{f}$, assume the inequality $f^{\Delta|ME}(\delta \,|\, \mu, \varepsilon) \leq \bar{f}$ holds for any $(\mu, \varepsilon) \in X$ and any δ between 0 and $\varepsilon(\nu - \mu)$. Then, the following bound holds:*

$$\left|\tau^e(X) - \tau(X)\right| \;\leq\; \bar{f}\,\sqrt{\Pr\{(M, E) \in X\}}\,\sqrt{\mathbb{E}\left[E^2\,(\nu - M)^2\right]}.$$

Proof. Using the assumed inequality, the desired bound is almost immediate:

$$\begin{aligned}
\left|\tau^e(X) - \tau(X)\right| &\leq \iint_X \left| \int_0^{\varepsilon(\nu-\mu)} \bar{f}\, d\delta \right| dF^{ME}(\mu,\varepsilon), \\
&= \bar{f} \iint_X \varepsilon\,|\nu-\mu|\, dF^{ME}(\mu,\varepsilon), \\
&= \bar{f} \iint_{\Omega^{ME}} \mathbb{1}_{\{(\mu,\varepsilon)\in X\}}\, \varepsilon\,|\nu-\mu|\, dF^{ME}(\mu,\varepsilon), \\
&\leq \bar{f} \sqrt{\iint_{\Omega^{ME}} \mathbb{1}^2_{\{(\mu,\varepsilon)\in X\}}\, dF^{ME}(\mu,\varepsilon)} \sqrt{\iint_{\Omega^{ME}} \varepsilon^2\,(\nu-\mu)^2\, dF^{ME}(\mu,\varepsilon)}, \\
&= \bar{f}\, \sqrt{\Pr\{(M,E)\in X\}}\, \sqrt{\mathbb{E}\left[E^2\,(\nu-M)^2\right]}.
\end{aligned}$$

The first inequality is assumed; the second is the Cauchy–Schwarz inequality. □

C.2.f.iv Vertical asymptotes in the threshold density.

In Example C.2, it was only at the endpoints of the domain of Θ that the density and its derivatives were both unbounded. Generalizing, we are most worried about vertical asymptotes cropping up in the threshold density at the endpoints of (a,b).

Consider $f^{\Theta|ME}(\theta\,|\,\mu,\varepsilon)$ having an asymptote as the threshold approaches a (or b). Necessarily, that density would be decreasing (increasing) for thresholds in some neighborhood of a (b). The monotonicity condition of the next result captures this intuition.

Result C.6. *Fix a set $X \subseteq \Omega^{ME} \setminus A$, and assume the inequality*

$$f^{\Delta|ME}(\delta\,|\,\mu,\varepsilon) \leq f^{\Delta|ME}(0\,|\,\mu,\varepsilon)$$

holds for any $(\mu,\varepsilon) \in X$ and any δ between 0 and $\varepsilon\,(\nu-\mu)$. Then, the following bound holds:

$$\left|\tau^e(X) - \tau(X)\right| \;\leq\; f^{\Delta}(0)\, \sqrt{\Pr\{(M,E)\in X \,|\, \Delta=0\}}\, \sqrt{\mathbb{E}\left[E^2\,(\nu-M)^2\,\middle|\,\Delta=0\right]}.$$

Proof. Since $f^{\Delta|ME}(\delta\,|\,\mu,\varepsilon) \geq 0$, it follows that $f^{\Delta|ME}(0\,|\,\mu,\varepsilon) - f^{\Delta|ME}(\delta\,|\,\mu,\varepsilon) \leq f^{\Delta|ME}(0\,|\,\mu,\varepsilon)$. And the assumed inequality tells us that $\left|f^{\Delta|ME}(\delta\,|\,\mu,\varepsilon) - f^{\Delta|ME}(0\,|\,\mu,\varepsilon)\right| = f^{\Delta|ME}(0\,|\,\mu,\varepsilon) - f^{\Delta|ME}(\delta\,|\,\mu,\varepsilon)$. Hence, we have shown that

$$\left|f^{\Delta|ME}(0\,|\,\mu,\varepsilon) - f^{\Delta|ME}(\delta\,|\,\mu,\varepsilon)\right| \leq f^{\Delta|ME}(0\,|\,\mu,\varepsilon).$$

Now, we can write

$$
\begin{aligned}
&\left|\tau^e(X) - \tau(X)\right| \\
&\quad \leq \iint_X \left| \int_0^{\varepsilon(\nu-\mu)} \left| f^{\Delta|ME}\left(\delta \,|\, \mu, \varepsilon\right) - f^{\Delta|ME}\left(0 \,|\, \mu, \varepsilon\right) \right| d\delta \right| \, dF^{ME}(\mu, \varepsilon), \\
&\quad \leq \iint_X \left| \int_0^{\varepsilon(\nu-\mu)} f^{\Delta|ME}\left(0 \,|\, \mu, \varepsilon\right) d\delta \right| \, dF^{ME}(\mu, \varepsilon), \\
&\quad = \iint_X \varepsilon \left|\nu - \mu\right| \, f^{\Delta|ME}\left(0 \,|\, \mu, \varepsilon\right) \, dF^{ME}(\mu, \varepsilon), \\
&\quad = f^{\Delta}(0) \iint_X \varepsilon \left|\nu - \mu\right| \, dF^{ME|\Delta}\left(\mu, \varepsilon \,|\, 0\right), \\
&\quad = f^{\Delta}(0) \iint_{\Omega^{ME}} \mathbb{1}_{\{(\mu,\varepsilon) \in X\}} \, \varepsilon \left|\nu - \mu\right| \, dF^{ME|\Delta}\left(\mu, \varepsilon \,|\, 0\right), \\
&\quad \leq f^{\Delta}(0) \sqrt{\iint_{\Omega^{ME}} \mathbb{1}^2_{\{(\mu,\varepsilon) \in X\}} \, dF^{ME|\Delta}\left(\mu, \varepsilon \,|\, 0\right)} \sqrt{\iint_{\Omega^{ME}} \varepsilon^2 \left(\nu - \mu\right)^2 \, dF^{ME|\Delta}\left(\mu, \varepsilon \,|\, 0\right)}, \\
&\quad = f^{\Delta}(0) \sqrt{\Pr\left\{(M, E) \in X \,|\, \Delta = 0\right\}} \sqrt{\mathbb{E}\left[E^2 \left(\nu - M\right)^2 \,|\, \Delta = 0\right]}.
\end{aligned}
$$

The first inequality is obvious, the second comes from the inequality we derived above, and the third is the Cauchy–Schwarz inequality. □

C.2.g Bringing it together

To fully bound our error, we will need to partition Ω^{ME} into parts where Results C.3, C.4, C.5 and C.6 apply. A **standard partition** is an ordered quadruple,

$$\Big(\left(A, \bar{\delta}\right), \, \left(B, k\right), \, \left(C, \bar{f}, \mathcal{E}^C\right), \, \left(D, \mathcal{E}^D\right) \Big),$$

where

- the set $\Omega^{ME} \setminus (A \cup B \cup C \cup D)$ is of F^{ME}-measure zero;
- the set A is as defined in Equation C.25 with the given $\bar{\delta}$;
- the set B satisfies the conditions of Result C.4 with the given k;
- the set C satisfies the conditions of Result C.5 with $\Pr\{(M, E) \in C\}$ bounded above by $\mathcal{E}^C$; and
- the set D satisfies the conditions of Result C.6 with $\Pr\{(M, E) \in D \,|\, \Delta = 0\}$ bounded above by $\mathcal{E}^D$.

Now, we introduce our full bound.

Proposition C.2. *Let $\left(A, (B, k), \left(C, \bar{f}, \mathcal{E}^C\right), \left(D, \mathcal{E}^D\right)\right)$ be a standard partition. Then, the following bound holds:*

$$
\begin{aligned}
\left|\tau^e - \tau\right| \leq \; & \frac{\mathbb{E}\left[E^2 \left(\nu - M\right)^2\right]}{\bar{\delta}^2} + \frac{f^{\Delta}(0)}{\bar{\delta}} \mathbb{E}\left[E^2 \left(\nu - M\right)^2 \,|\, \Delta = 0\right] \\
& + \frac{k}{2} \mathbb{E}\left[E^2 \left(\nu - M\right)^2\right] + \bar{f} \sqrt{\mathcal{E}^C} \sqrt{\mathbb{E}\left[E^2 \left(\nu - M\right)^2\right]} \\
& + f^{\Delta}(0) \sqrt{\mathcal{E}^D} \sqrt{\mathbb{E}\left[E^2 \left(\nu - M\right)^2 \,|\, \Delta = 0\right]}
\end{aligned}
$$

Proof. Use the triangle inequality on the bounds from Results C.3, C.4, C.5, and C.6. □

To see how this works more concretely, let's construct a standard partition for Example C.2 with parameters $(p, q) = (3/2, 1/2)$. Further, let's set $\nu = 3/5$.

- $(A, \bar{\delta})$**:** The density is valid everywhere, so we can set $\bar{\delta}$ to whatever we like. We choose $\bar{\delta} = 1/8$ and use Equation C.25 to define A. For any $(\mu, \varepsilon) \in \Omega^{ME} \setminus A$ then, we know that $\varepsilon\,|\nu - \mu| \leq 1/8$.
- (B, k)**:** Pick any $\eta \in (0, 1/8)$, and let $B = \big([\eta, 1-\eta] \times \Omega^E\big) \setminus A$. For any $(\mu, \varepsilon) \in B$, we know that μ and $\mu + \varepsilon\,(\nu - \mu)$ (and all numbers between) are in $[\eta, 1-\eta]$. Looking at Figure C.5, on $[\eta, 1-\eta]$, the density is continuous and its slope is bounded, so we can construct a Lipschitz constant, k, in much the same way we did in Example C.1.
- $(C, \bar{f}, \mathcal{E}^C)$**:** Let $C = \big((0, \eta) \times \Omega^E\big) \setminus A$, and let $\bar{f} = f^{\Theta|ME}\big(\eta + 1/8 \,\big|\, \mu, \varepsilon\big)$. Looking at Figure C.5, on $(0, \eta + 1/8)$, the density is smaller than $\bar{f}$. Finally, note that $\Pr\{(M, E) \in C\} \leq \Pr\big\{M \in (0, \eta)\big\}$. So, we set $\mathcal{E}^C = \Pr\big\{M \in (0, \eta)\big\}$.
- $(D, \mathcal{E}^D)$**:** Let $D = \big((1-\eta, 1) \times \Omega^E\big) \setminus A$. The conditions of Result C.6 are met since the density is strictly increasing on $\big(1 - (\eta + \bar{\delta}), 1\big)$. Finally, note that $\Pr\{(M, E) \in D\} \leq \Pr\big\{M \in (1-\eta, 1)\big\}$. So, we set $\mathcal{E}^D = \Pr\big\{M \in (1-\eta, 1)\big\}$.

Points of interest from this standard partition are illustrated in Figure C.5; the partition itself is illustrated in Figure C.6. Given all of this, Proposition C.2 gives us a bound on the absolute error of our approximation.

What's more, by decreasing η, we can make $\mathcal{E}^C$ and $\mathcal{E}^D$ as small as we like. We call an experiment **regular** if there is some $\bar{\bar{f}}$ such that, for any $\mathcal{E} > 0$, we can construct a standard partition for which $\max\{\mathcal{E}^C, \mathcal{E}^D\} < \mathcal{E}$ and $\bar{f} \leq \bar{\bar{f}}$.

C.2.h Asymptotics

Formally, asymptotic analysis requires considering a sequence of experiments, $\big((\nu_n, F_n^{\Delta ME})\big)_{n=1}^{\infty}$. We capture dependence on n in a derived quantity with an n subscript; for example, $f_n^{\Delta}(0)$ represents the density of marginal agents in the nth experiment.

Now, to start, we will introduce our set of baseline asymptotic assumptions. The first two merely adjust Assumptions C.3 and C.7 to the asymptotic context. The third assumes that there are in fact marginal agents, while the last asserts that all experiments in the asymptotic sequence are regular in a coherent way.

Assumption C.8. *The following are true:*

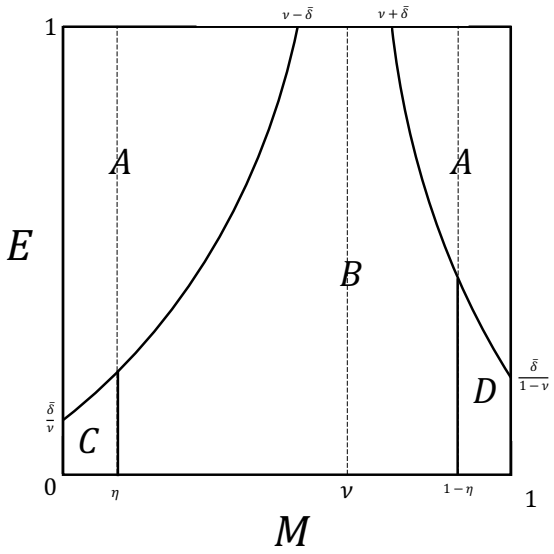


Figure C.6: Standard partition from the end of Section C.2.g

a) *Assumption C.3 holds for all n uniformly; that is, the sequence $(\bar{\delta}_n)$ is uniformly bounded away from zero.*
b) *For all n, Assumption C.7 holds.*
c) *The sequence $\left(f_n^{\Delta}(0)\right)$ is uniformly bounded above and uniformly bounded away from zero.*
d) *The sequence of experiments is uniformly regular. That is, there is an $\bar{\bar{f}}$ such that, for any $\mathcal{E}>0$, for each n, there is a standard partition,*

$$\left(\left(A_n, \bar{\delta}_n\right), \left(B_n, k_n\right), \left(C_n, \bar{f}_n, \mathcal{E}_n^C\right), \left(D_n, \mathcal{E}_n^D\right)\right),$$

such that $\max\{\mathcal{E}_n^C, \mathcal{E}_n^D\} < \mathcal{E}$, the sequence (k_n) is uniformly bounded above, and the sequence $(\bar{f}_n)$ is uniformly bounded above by $\bar{\bar{f}}$.

From here, we will show that the absolute error converges to zero and then discuss why we must shift our focus to the relative error.

C.2.h.i Absolute Error

To capture the intuition of a nudge, we assume that updates get small in our asymptotic sequence. The following assumption meets our technical needs.

Assumption C.9. *The following limits hold:*

$$\lim_{n\to\infty} \mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right] = 0 = \lim_{n\to\infty} \mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2 \middle| \Delta = 0\right].$$

Jensen's inequality then immediately tells us that

Result C.7. *The following limits hold:*

$$\lim_{n\to\infty} \mathbb{E}_n\left[E\left(\nu_n - M\right)\right] \;=\; 0 \;=\; \lim_{n\to\infty} \mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\, \Delta = 0\right].$$

It follows almost immediately then that the absolute error goes to zero.

Proposition C.3. *The following limit holds:*

$$\lim_{n\to\infty} \left|\tau_n^e - \tau_n\right| \;=\; 0.$$

Proof. Consider the bound from Proposition C.2. Pick any $\mathcal{E} > 0$. The elements of Assumption C.8 tell us that $1/\delta_n^2$, $f_n^\Delta(0)/\delta_n$, k_n, $\bar{f}_n$, and $f_n^\Delta(0)$ are all bounded above. And $\mathcal{E}_n^C$ and $\mathcal{E}_n^D$ are both less than $\mathcal{E}$. Hence, Assumption C.9 establishes our final, desired result. □

At first glance, this result seems sufficient. But, it implies that

Result C.8. *The following limits hold:*

$$\lim_{n\to\infty} \tau_n \;=\; 0 \;=\; \lim_{n\to\infty} \tau_n^e.$$

Proof. Consider the formula for τ_n, that is, $f_n^\Delta(0)\, \mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\, \Delta = 0\right]$. By Assumption C.8 and Result C.7, this approaches zero, and we've established the first desired limit.

Now, by the triangle inequality, $|\tau_n^e| \leq |\tau_n| + |\tau_n^e - \tau_n|$. Using the limit just established and Proposition C.3, this becomes $\lim_{n\to\infty} |\tau_n^e| \leq 0$, which establishes the second desired limit. □

This result tells us that, asymptotically, τ_n^e and the constant 0 have the same absolute error (i.e., zero). But 0 isn't a particularly compelling approximation for the treatment effect of a nudge!

To see where τ_n outperforms 0, we will need to consider *relative* error. In the next section, we will show that τ_n and τ_n^e don't just become infinitesimally close, but that their ratio approaches one. In other words, we will show that they are **asymptotically equivalent**.

C.2.h.ii Relative error

Before considering relative error, we must make a few more assumptions about the asymptotic sequence.

Assumption C.10. *The following regularity conditions hold:*

a) *The sequence*

$$\left(\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\left(\mathbb{E}_n\left[E\left(\nu_n - M\right)\right]\right)^2}\right)_{n=1}^{\infty}$$

is defined and uniformly bounded above and below by strictly positive constants.[6a]

b) *For $j \in \{1, 2\}$, the sequence*

$$\left(\left|\frac{\mathbb{E}_n\left[E^j\left(\nu_n - M\right)^j\right]}{\mathbb{E}_n\left[E^j\left(\nu_n - M\right)^j \middle| \Delta = 0\right]}\right|\right)_{n=1}^{\infty}$$

is defined and uniformly bounded above and below by strictly positive constants.

Part a) ensures that $\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]$ and $\left(\mathbb{E}_n\left[E\left(\nu_n - M\right)\right]\right)^2$ shrink at about the same rate. Part b) ensures that marginal agents aren't too different from the general population.

Now, we can use the bound from Proposition C.2 to show that the ratio of τ_n^e / τ_n converges to unity.

Theorem C.1. *The ratio τ_n^e / τ_n is defined for all n; moreover, the following limit holds:*

$$\lim_{n \to \infty} \left|\tau_n^e / \tau_n - 1\right| = 0.$$

Proof. For the ratio to be defined, we need τ_n to always be non-zero. This is true because Assumption C.8 tells us that $f_n^{\Delta}(0)$ is always positive, while Assumption C.10 tells us that $\mathbb{E}_n\left[E\left(\nu_n - M\right) \middle| \Delta = 0\right]$ is always non-zero.

Hence, we can divide anything through by τ_n in our asymptotic. Doing so with the bound from Proposition C.2 gives

$$\begin{aligned}\left|\tau_n^e / \tau_n - 1\right| \leq\ & \underbrace{\frac{1}{\bar{\delta}_n^2 \, f_n^{\Delta}(0)} \frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right) \middle| \Delta = 0\right]\right|}}_{\text{Term A1}} + \underbrace{\frac{1}{\bar{\delta}} \frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2 \middle| \Delta = 0\right]}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right) \middle| \Delta = 0\right]\right|}}_{\text{Term A2}} \\ & + \underbrace{\frac{k_n}{2 \, f_n^{\Delta}(0)} \frac{\mathbb{E}_n\left[E^2\left(\nu - M\right)^2\right]}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right) \middle| \Delta = 0\right]\right|}}_{\text{Term B}} + \underbrace{\frac{\bar{f}_n}{f_n^{\Delta}(0)} \sqrt{\mathcal{E}_n^C} \frac{\sqrt{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right) \middle| \Delta = 0\right]\right|}}_{\text{Term C}} \\ & + \underbrace{\sqrt{\mathcal{E}_n^D} \frac{\sqrt{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2 \middle| \Delta = 0\right]}}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right) \middle| \Delta = 0\right]\right|}}_{\text{Term D}}.\end{aligned}$$

The terms are named to correspond the part of the standard partition they correspond to. We consider the convergence of each term individually.

6a. The lower bound here isn't really an assumption, as Jensen's inequality ensures that the sequence is bounded below by 1.

Term A1. By Assumption C.8, Term A1 converges to zero if

$$\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right|}$$

converges to zero. To see that it does, note that

$$\begin{aligned}
&\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right|} \\
&\quad = \underbrace{\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\left(\mathbb{E}_n\left[E\left(\nu_n - M\right)\right]\right)^2}}_{\text{Term A1-i}} \underbrace{\left(\frac{\mathbb{E}_n\left[E\left(\nu_n - M\right)\right]}{\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]}\right)^2}_{\text{Term A1-ii}} \underbrace{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right|}_{\text{Term A1-iii}}.
\end{aligned}$$

Terms A1-i and A1-ii are bounded above and below by Assumption C.10. Term A1-iii goes to zero by Result C.7. So, Term A1 converges to zero.

Term A2. By Assumption C.8, Term A2 converges to zero if

$$\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\middle|\Delta = 0\right]}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right|}$$

converges to zero. To see that it does, note that

$$\begin{aligned}
&\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\middle|\Delta = 0\right]}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right|} \\
&\quad = \underbrace{\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\middle|\Delta = 0\right]}{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}}_{\text{Term A2-i}} \underbrace{\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\left(\mathbb{E}_n\left[E\left(\nu_n - M\right)\right]\right)^2}}_{\text{Term A2-ii}} \underbrace{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right)\right]\right|}_{\text{Term A2-iii}}.
\end{aligned}$$

Terms A2-i and A2-ii are bounded above and below by Assumption C.10. Term A2-iii goes to zero by Result C.7. So, Term A2 converges to zero.

Term B. By Assumption C.8, Term B converges to zero if

$$\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right|}$$

converges to zero. To see that it does, note that

$$\begin{aligned}
&\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right|} \\
&= \underbrace{\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\middle|\Delta = 0\right]}}_{\text{Term B-i}} \underbrace{\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\middle|\Delta = 0\right]}{\left(\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right)^2}}_{\text{Term B-ii}} \underbrace{\left|\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right|}_{\text{Term B-iii}}.
\end{aligned}$$

Terms B-i and B-ii are bounded above and below by Assumption C.10. Term B-iii goes to zero by Result C.7. So, Term B converges to zero.

Term C. Term C can be rewritten as

$$\begin{aligned}
&\frac{\bar{f}_n}{f_n^{\Delta}(0)}\sqrt{\mathcal{E}_n^C}\sqrt{\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\left(\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right)^2}} \\
&\quad = \frac{\bar{f}_n}{f_n^{\Delta}(0)}\sqrt{\mathcal{E}_n^C}\sqrt{\underbrace{\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\right]}{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\middle|\Delta = 0\right]}}_{\text{Term C-i}} \underbrace{\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2\middle|\Delta = 0\right]}{\left(\mathbb{E}_n\left[E\left(\nu_n - M\right)\middle|\Delta = 0\right]\right)^2}}_{\text{Term C-ii}}}.
\end{aligned}$$

Terms C-i and C-ii are bounded above by Assumption C.10. Hence the radical on the right-hand side of the previous equation is bounded above. Further, by Assumption C.8, $\bar{f}_n \Big/ f_n^{\Delta}(0)$ is bounded above. Hence, Term C is bounded above by the product of some positive constant, K^C, and $\sqrt{\mathcal{E}_n^C}$.

Term D. Term D can be rewritten as

$$\sqrt{\mathcal{E}_n^D} \sqrt{\frac{\mathbb{E}_n\left[E^2\left(\nu_n - M\right)^2 \middle| \Delta = 0\right]}{\left(\mathbb{E}_n\left[E\left(\nu_n - M\right) \middle| \Delta = 0\right]\right)^2}}.$$

The rightmost radical is bounded above by Assumption C.10. Hence,Term D is bounded above by the product of some positive constant, K^D, and $\sqrt{\mathcal{E}_n^D}$.

Combining the bounds for Terms A–D, we get that

$$\lim_{n\to\infty} \left|\tau_n^e / \tau_n - 1\right| \leq \left(K^C + K^D\right) \sqrt{\max\{\mathcal{E}_n^C, \mathcal{E}_n^D\}}.$$

In other words, the relative error, $\left|\tau_n^e / \tau_n - 1\right|$, is bounded by the product of a constant, $K^C + K^D$, and something, by Assumption C.8, we can make arbitrarily small. This establishes our desired result. □

That is, τ_n is non-zero, finite, and asymptotically equivalent to τ_n^e. Based on this result, going forward, when we refer to the treatment effect, we refer to τ.

C.3 Literature-level model

C.3.a The big idea

Variations in the baseline and treatment effect are driven by moving the nudge signal and the distribution of (Θ, M, E). To capture the first idea, at the literature level, we model the nudge signal as the realization of a random variable, N. To capture the second idea, we assume the distribution of (Θ, M, E) is changed by some random vector of **literature-level parameters**.

C.3.b Simplifying the experiment-level model

Before elaborating on the approach laid out in the previous section, it is helpful to introduce two changes to the experiment-level model. The first makes it easier to compare experiments; the second simplifies what follows.

C.3.b.i Holding the belief distribution fixed across experiments

For any strictly increasing, continuous transformation, H, consider setting

$$\begin{aligned}
\widetilde{M} &= H(M), \\
\widetilde{\Theta} &= H(\Theta), \text{ and} \\
\widetilde{\nu} &= H(\nu).
\end{aligned}$$

Further, set

$$\widetilde{E} \;=\; \begin{cases} \dfrac{H\big(M+E(\nu-M)\big)-\widetilde{M}}{\widetilde{\nu}-\widetilde{M}} & \text{if } \widetilde{\nu}\neq\widetilde{M}, \\ E & \text{if } \widetilde{\nu}=\widetilde{M}.^{7a} \end{cases}$$

Clearly, $M \geq \Theta$ if and only if $\widetilde{M} \geq \widetilde{\Theta}$, and $M + E(\nu - M) \geq \Theta$ if and only if $\widetilde{M} + \widetilde{E}\big(\widetilde{\nu} - \widetilde{M}\big) \geq \widetilde{\Theta}$. In other words, the model of the individual is invariant under strictly increasing, continuous transformations.

Now, take the distribution function for the standard normal, Φ. If we assume F^M is strictly increasing and continuous (i.e., has no atoms and is of full support), then the function $\Phi^{-1} \circ F^M$ is also strictly increasing and continuous; furthermore, if we set H equal to it, the transformation above leaves priors distributed according to the standard normal and leaves thresholds, posteriors, and the nudge signal to be interpreted as z-scores relative to the prior distribution.

Going forward we will always consider experiments after they have been transformed in the way just described. So, while all experiments have unit-normal priors, each is potentially different in its nudge signal, ν, and its conditional-on-prior joint-distribution of thresholds and update-strengths, $F^{\Theta E|M}$. Summarizing:

Assumption C.11 (Fixed belief distribution)**.** *For all experiments, the distribution function for beliefs, F^M, is that of the standard normal, Φ.*

Note this means that the interval (a, b) is now the real line for all experiments.

C.3.b.ii Independence of preferences and the updating process

Intuitively, thresholds summarize preferences, while the prior and update strength summarize information processing. As such, it is sensible to assume, for a given experiment, that thresholds are independent of priors and update strengths.

Assumption C.12 (Independent thresholds)**.** *For any experiment, the threshold, Θ, is independent of the prior and update strength, (M, E).*

Given this assumption, experiments vary in nudge signal, ν, threshold distribution, F^Θ, and conditional-on-prior update-strength distribution, $F^{E|M}$. Note that when combined with Assumption C.3, Assumption C.12 means that the threshold has a continuous density, f^Θ, over the entire real line.

7a. Our $\widetilde{\nu} = \widetilde{M}$ definition of $\widetilde{E}$ has the nice property that, when H is differentiable at ν, it equals the limit of the $\widetilde{\nu} \neq \widetilde{M}$ definition as M approaches ν.

C.3.c The literature's data-generating process

Given the reparametrization of the previous section, we set the outcome space for the random variable N to the real line. Moving on to the literature-level parameters, without loss of generality, we assume one of them, B, is the baseline, and we set its outcome space to be the open unit interval.[8a] Aside from B, we assume an additional random vector of literature-level parameters, Λ, whose outcome space is some subset, $\mathscr{L}$, of a finite-dimensional Euclidean space. We call Λ the **literature-level noise**. When the realization of (B, Λ) is (β, λ), the distributions that describe (Θ, M, E) are $F^{\Theta}_{\beta,\lambda}$ and $F^{E|M}_{\beta,\lambda}$.

In sum, a given experiment can be seen as a realization of the random vector (N, B, Λ), whose outcome space is $\mathbb{R} \times (0,1) \times \mathscr{L}$. The distribution of this random vector can then be described by some measure, $G^{NB\Lambda}$. To capture the idea that Λ really is noise, we will also make an independence assumption.

Assumption C.13 (Independence of the literature-level noise)**.** *The random vector Λ is independent of the random vector (N, B).*

Given this assumption, the literature can be summarized by two measures: G^{NB} and G^{Λ}. We will also assume that the conditional-on-baseline first moment of N exists and is uniformly bounded.

Assumption C.14. *There exists some $\bar{\nu} > 0$ such that, for any β on the open unit interval, $\mathbb{E}\left[|N| \,\middle|\, B = \beta\right] \leq \bar{\nu}$.*

C.3.d The baseline and treatment-effect integrals

Given the setup just discussed, when the realization of (B, Λ) is (β, λ), the equation

$$\beta = \int_{-\infty}^{\infty} F^{\Theta}_{\beta,\lambda}(\mu)\, d\Phi(\mu) \tag{C.27}$$

must be satisfied. Now, it is clear why we defined the outcome space of B to omit zero and one: such baselines would require the integrand of the previous equation to be 0 everywhere or 1 everywhere, respectively. No distribution function can satisfy such a requirement. That said, it is still possible for valid distribution functions to achieve baselines arbitrarily close to zero and one.

Moving on from the baseline, when the realization of (N, B, Λ) is (ν, β, λ), we can write the treatment effect as

$$\tau(\nu, \beta, \lambda) = \int_{-\infty}^{\infty} \mathbb{E}_{\beta,\lambda}\left[E \,\middle|\, M = \mu\right] (\nu - \mu)\, f^{\Theta}_{\beta,\lambda}(\mu)\, d\Phi(\mu), \tag{C.28}$$

8a. The reason the outcome space for B doesn't include 0 and 1 is technical in nature and will be clarified shortly.

where

$$\mathbb{E}_{\beta,\lambda}\left[E \,\middle|\, M=\mu\right] = \int_0^1 \varepsilon \, dF_{\beta,\lambda}^{E|M}\left(\varepsilon \,\middle|\, \mu\right).$$

The (β, λ) subscript on the expectation indicates that it is a function of those realizations.

In what follows, we build on the foundation of Equations C.27 and C.28 to learn how two important quantities vary with the baseline: the *probability* that the treatment effect is positive and the *expectation* of the treatment effect.

C.3.e What drives variation in the baseline?

Looking at Equation C.27, it is clear that β must shift thresholds down (and hence *increase* the distribution function $F_{\beta,\lambda}^{\Theta}$) in some sense. To add structure to this intuition, we assume the parameter β moves the threshold distribution in the likelihood-ratio sense. That is,

Assumption C.15 (β shifts Θ down in the likelihood-ratio sense)**.** *For any realization, λ, of Λ; any two real numbers θ' and θ; and any two baselines, β' and β, on $(0,1)$; if $\theta' > \theta$ and $\beta' > \beta$, then the inequality*

$$f_{\beta,\lambda}^{\Theta}(\theta')\, f_{\beta',\lambda}^{\Theta}(\theta) > f_{\beta',\lambda}^{\Theta}(\theta')\, f_{\beta,\lambda}^{\Theta}(\theta)$$

is satisfied.

This assumption implies that increasing β decreases the threshold distribution in the first-order-stochastic sense. That is,

Result C.9 (β shifts Θ down in the first-order stochastic sense)**.** *For any realization, λ, of Λ, and any θ on the real line, the distribution function $F_{\beta,\lambda}^{\Theta}(\theta)$ is strictly increasing in β.*

Proof. Take (θ', θ'') satisfying $\theta' < \theta < \theta''$ and $\beta' > \beta$. Assumption C.15 tells us that

$$f_{\beta,\lambda}^{\Theta}(\theta'')\, f_{\beta',\lambda}^{\Theta}(\theta') > f_{\beta',\lambda}^{\Theta}(\theta'')\, f_{\beta,\lambda}^{\Theta}(\theta').$$

Integrating over all $\theta'' > \theta$ and $\theta' < \theta$ yields

$$F_{\beta,\lambda}^{\Theta}(\theta)\left(1 - F_{\beta',\lambda}^{\Theta}(\theta)\right) < \left(1 - F_{\beta,\lambda}^{\Theta}(\theta)\right) F_{\beta',\lambda}^{\Theta}(\theta).$$

Adding $F_{\beta,\lambda}^{\Theta}(\theta)\, F_{\beta',\lambda}^{\Theta}(\theta)$ to both sides yields the desired result. □

Corollary 1. *For any realization, λ, of Λ, the baseline integral,*

$$\int_{-\infty}^{\infty} F_{\beta,\lambda}^{\Theta}(\mu)\, d\Phi(\mu),$$

is strictly increasing in β.

The corollary shows that Assumption C.15 is consistent with Equation C.27. Assumption C.15 also means that the threshold density is non-zero everywhere.

Result C.10 (Threshold density is non-zero)**.** *For any realization, λ, of Λ, β' on $(0,1)$, and θ on the real line, the threshold density obeys the inequality $f^{\Theta}_{\beta',\lambda}(\theta) > 0$.*

Proof. Pick $\beta < \beta'$ and $\theta' > \theta$. If $f^{\Theta}_{\beta',\lambda}(\theta) = 0$, Assumption C.15 reads

$$0 > f^{\Theta}_{\beta',\lambda}(\theta')\, f^{\Theta}_{\beta,\lambda}(\theta),$$

which cannot hold, since densities are non-negative. □

C.3.f What drives variation in the sign of the treatment effect?

Recall the expression for the treatment effect in Equation C.24, which shows that, when the realization of (N, B, Λ) is (ν, β, λ), the treatment effect is strictly positive if and only if

$$f^{\Delta}_{\beta,\lambda}(0)\ \mathbb{E}_{\beta,\lambda}\left[E \,\middle|\, \Delta = 0\right] \ > \ 0 \tag{C.29}$$

and

$$\nu \ > \ \frac{\mathbb{E}_{\beta,\lambda}\left[E\,M \,\middle|\, \Delta = 0\right]}{\mathbb{E}_{\beta,\lambda}\left[E \,\middle|\, \Delta = 0\right]}. \tag{C.30}$$

A basic assumption ensures the first inequality is satisfied. So, variation in the sign of the treatment effect is ultimately driven by variation in the two sides of the second inequality.

The first inequality

We can write the right-hand side of the inequality in Equation C.29 more explicitly as

$$f^{\Delta}_{\beta,\lambda}(0)\ \mathbb{E}_{\beta,\lambda}\left[E \,\middle|\, \Delta = 0\right] = \int_{-\infty}^{\infty} \mathbb{E}_{\beta,\lambda}\left[E \,\middle|\, M = \mu\right] f^{\Theta}_{\beta,\lambda}(\mu)\, d\Phi(\mu). \tag{C.31}$$

Result C.10 tells us that $f^{\Theta}_{\beta,\lambda}(\mu)$ is strictly positive, so for the entire integral to be positive, we just have to make an assumption about $\mathbb{E}_{\beta,\lambda}\left[E \,\middle|\, M = \mu\right]$.

Assumption C.16 (Agents update)**.** *The conditional-on-prior expected update strength, $\mathbb{E}_{\beta,\lambda}\left[E \,\middle|\, M = \mu\right]$, is strictly positive for all μ, regardless of the realization, (β, λ), of (B, Λ).*

This assumption is rather uncontroversial; it states that the sub-population with almost any prior doesn't consist *entirely* of agents that ignore the signal. But, it does give us the following useful result.

Result C.11. *The inequality in Equation C.29 holds.*

Now, we move on to the two sides of the inequality in Equation C.30.

Marginal priors

We can write the right-hand side of the inequality in Equation C.30 more explicitly as

$$\frac{\mathbb{E}_{\beta,\lambda}\left[E\,M\,\middle|\,\Delta=0\right]}{\mathbb{E}_{\beta,\lambda}\left[E\,\middle|\,\Delta=0\right]}=\frac{\int_{-\infty}^{\infty}\mu\,\mathbb{E}_{\beta,\lambda}\left[E\,\middle|\,M=\mu\right]f^{\Theta}_{\beta,\lambda}(\mu)\,d\Phi(\mu)}{\int_{-\infty}^{\infty}\mathbb{E}_{\beta,\lambda}\left[E\,\middle|\,M=\mu\right]f^{\Theta}_{\beta,\lambda}(\mu)\,d\Phi(\mu)}. \tag{C.32}$$

Instead of thinking of this as the ratio of two expectations, we can think of it as one: the expected marginal prior *weighted by update strength*. Looking back at Equation C.32, this is a simple first moment with respect to the density

$$\psi_{\beta,\lambda}(\mu)=\frac{\mathbb{E}_{\beta,\lambda}\left[E\,\middle|\,M=\mu\right]f^{\Theta}_{\beta,\lambda}(\mu)\,\varphi(\mu)}{\int_{-\infty}^{\infty}\mathbb{E}_{\beta,\lambda}\left[E\,\middle|\,M=\tilde{\mu}\right]f^{\Theta}_{\beta,\lambda}(\tilde{\mu})\,\varphi(\tilde{\mu})\,d\tilde{\mu}},$$

whose likelihood ratio can be written

$$\frac{\psi_{\beta,\lambda}(\mu')}{\psi_{\beta,\lambda}(\mu)}=\frac{f^{\Theta}_{\beta,\lambda}(\mu')}{f^{\Theta}_{\beta,\lambda}(\mu)}\,\frac{\mathbb{E}_{\beta,\lambda}\left[E\,\middle|\,M=\mu'\right]}{\mathbb{E}_{\beta,\lambda}\left[E\,\middle|\,M=\mu\right]}.$$

So long as increasing the baseline doesn't increase the relative update strengths of agents with high priors (too much), this density will inherit the maximum-likelihood property from Assumption C.15. We assume this is the case.

Assumption C.17 (β shifts marginal agents' update-weighted priors down in the likelihood-ratio sense)**.** *For any realization, λ, of Λ; any two real numbers μ' and μ; and any two baselines, β and β', on $(0,1)$; if $\mu'>\mu$ and $\beta'>\beta$, then the inequality*

$$\psi_{\beta,\lambda}(\mu')\,\psi_{\beta',\lambda}(\mu)\;>\;\psi_{\beta',\lambda}(\mu')\,\psi_{\beta,\lambda}(\mu)$$

is satisfied.

This immediately implies that the distribution function that corresponds to $\psi_{\beta,\lambda}$—i.e., $\Psi_{\beta,\lambda}(\mu)\equiv\int_{-\infty}^{\mu}\psi_{\beta,\lambda}(\tilde{\mu})\,d\tilde{\mu}$—is increasing in β.

Result C.12 (β shifts marginal agents' update-weighted priors down in the first-order stochastic sense)**.** *For any realization, λ, of Λ, any real number μ, the distribution function $\Psi_{\beta,\lambda}(\mu)$ is strictly increasing in the baseline, β.*

Proof. This proof is almost identical to that of Result C.9. □

From here it is clear that the expectation of the update-weighted prior among the marginal is decreasing in baseline.

Corollary 2 (β shifts the expectation of marginal agents' update-weighted priors down)**.** *For any realization, λ, of Λ, the expectation of update-weighted priors among the marginal, $\mathbb{E}_{\lambda,\beta}\left[E\,M \middle| \Delta = 0\right] / \mathbb{E}_{\lambda,\beta}\left[E \middle| \Delta = 0\right]$, is strictly decreasing in β.*

Signal

Given the results of the previous section, we see that increasing β can only switch the sign of the treatment effect from negative to positive, so long as the signal, N, isn't too negatively correlated with the baseline. To simplify things, we make the following assumption.

Assumption C.18. *The conditional-on-baseline signal distribution is increasing in the first-order stochastic sense with baseline; that is, for any real ν, the function $G^{N|B}\left(\nu \middle| \beta\right)$ is weakly decreasing in β.*

We will also assume that, for any baseline, the conditional-on-baseline nudge signal distribution has full support. This is a technical convenience: we could proceed without it at the expense of making the statements of subsequent theorems a but more complicated.

Assumption C.19. *For any baseline, β, the conditional-on-baseline signal distribution, $G^{N|B}\left(\nu \middle| \beta\right)$ is strictly increasing in ν.*

Now, we are prepared to show how the probability of a positive treatment effect varies with the baseline.

The probability of a positive treatment effect

Looking back to Equation C.30, we see that, conditional on baseline, the probability of a strictly positive treatment effect is given by

$$\Pr\{\tau(B,N,\Lambda) > 0 \,|\, B = \beta\} = 1 - \mathbb{E}\left[G^{N|B}\left(\frac{\mathbb{E}_{B,\Lambda}\left[E\,M \middle| \Delta = 0\right]}{\mathbb{E}_{B,\Lambda}\left[E \middle| \Delta = 0\right]} \middle| B = \beta\right)\right],$$

where the expectation is over the literature-level noise, Λ. Given the assumptions from the previous subsection, it is easy to see that the probability of a positive treatment effect is increasing in the baseline.

Theorem C.2. *The function $\Pr\{\tau(B,N,\Lambda) > 0 \,|\, B = \beta\}$ is strictly increasing in β.*

Proof. Take any realization, λ, of Λ, and any two baselines on the open unit interval that satisfy $\beta' > \beta$. By Corollary 2,

$$\mathbb{E}_{\lambda,\beta'}\left[E\,M \middle| \Delta = 0\right] / \mathbb{E}_{\lambda,\beta'}\left[E \middle| \Delta = 0\right] \;<\; \mathbb{E}_{\lambda,\beta}\left[E\,M \middle| \Delta = 0\right] / \mathbb{E}_{\lambda,\beta}\left[E \middle| \Delta = 0\right].$$

Since distribution functions are strictly increasing (by Assumption C.19), this means that

$$G^{N|B}\left(\mathbb{E}_{\lambda,\beta'}\left[EM\middle|\Delta=0\right]/\mathbb{E}_{\lambda,\beta'}\left[E\middle|\Delta=0\right]\middle|B=\beta\right) < G^{N|B}\left(\mathbb{E}_{\lambda,\beta}\left[EM\middle|\Delta=0\right]/\mathbb{E}_{\lambda,\beta}\left[E\middle|\Delta=0\right]\middle|B=\beta\right). \quad (\star)$$

By Assumption C.18, we also know

$$G^{N|B}\left(\mathbb{E}_{\lambda,\beta'}\left[EM\middle|\Delta=0\right]/\mathbb{E}_{\lambda,\beta'}\left[E\middle|\Delta=0\right]\middle|B=\beta'\right) \leq G^{N|B}\left(\mathbb{E}_{\lambda,\beta'}\left[EM\middle|\Delta=0\right]/\mathbb{E}_{\lambda,\beta'}\left[E\middle|\Delta=0\right]\middle|B=\beta\right).$$

Putting the previous two inequalities together yields

$$G^{N|B}\left(\mathbb{E}_{\lambda,B}\left[EM\middle|\Delta=0\right]/\mathbb{E}_{\lambda,B}\left[E\middle|\Delta=0\right]\middle|B=\beta'\right) < G^{N|B}\left(\mathbb{E}_{\lambda,B}\left[EM\middle|\Delta=0\right]/\mathbb{E}_{\lambda,B}\left[E\middle|\Delta=0\right]\middle|B=\beta\right).$$

Taking the expectation over realizations of Λ and subtracting from one yields

$$\Pr\{\tau(B,N,\Lambda)>0\,|\,B=\beta'\} > \Pr\{\tau(B,N,\Lambda)>0\,|\,B=\beta\},$$

as desired. Note that this inequality would be weak if $G^{N|B}(\cdot\,|\,\beta)$ were constant from between $\mathbb{E}_{\lambda,\beta'}\left[EM\middle|\Delta=0\right]/\mathbb{E}_{\lambda,\beta'}\left[E\middle|\Delta=0\right]$ and $\mathbb{E}_{\lambda,\beta}\left[EM\middle|\Delta=0\right]/\mathbb{E}_{\lambda,\beta}\left[E\middle|\Delta=0\right]$, as it would make the inequality marked with $(\star)$ weak. The charm of Assumption C.19 is that it allows us to side-step this detail. □

C.3.g The conditional-on-baseline expected treatment effect

In this section, we will show that the conditional-on-baseline expected treatment effect, $\mathbb{E}\left[\tau(N,B,\Lambda)\middle|B=\beta\right]$, approaches zero as the baseline approaches 0 or 1 and that it is single crossing from below. To do so, we must first discuss averaging out the literature-level noise.

Averaging out the literature-level noise

If we take the conditional-on-baseline expectation of Equation C.27, Fubini's theorem lets us write

$$\beta = \int_{-\infty}^{\infty} F_{\beta}^{\Theta}(\mu)\, d\Phi(\mu),$$

where

$$F_{\beta}^{\Theta}(\theta) \equiv \mathbb{E}\left[F_{\beta,\Lambda}^{\Theta}(\theta)\right].$$

If we further define

$$f_{\beta}^{\Theta}(\theta) \equiv \mathbb{E}\left[f_{\beta,\Lambda}^{\Theta}(\theta)\right],$$

we can show the following result.

Result C.13. *For any $\beta \in (0,1)$, the distribution F_{β}^{Θ} represents a probability measure on the real line with density f_{β}^{Θ}.*

Proof. Fubini's theorem shows that, for any real θ, $F^{\Theta}_{\beta}(\theta) = \int_{-\infty}^{\theta} f^{\Theta}_{\beta}(\theta')\, d\theta'$ and $\int_{-\infty}^{\infty} f^{\Theta}_{\beta}(\theta)\, d\theta = 1$. □

The probability measure represented by F^{Θ}_{β} is best thought of as a Λ-averaged version of the measure represented by $F^{\Theta}_{\Lambda,\beta}$. Going forward, it will prove useful to show that f^{Θ}_{β} is strictly positive.

Result C.14. *For any $\beta \in (0,1)$ and any real θ, the density $f^{\Theta}_{\beta}(\theta)$ is strictly positive.*

Proof. This is a straightforward consequence of Result C.10. □

Turning to the treatment effect, if we take the conditional-on-baseline expectation of Equation C.28, Fubini's theorem lets us write

$$\begin{aligned}\mathbb{E}\left[\tau(N,B,\Lambda)\,\middle|\,B=\beta\right] \\ = \int_{-\infty}^{\infty} \mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right] f^{\Theta}_{\beta}(\mu)\left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu\right) d\Phi(\mu),\end{aligned} \tag{C.33}$$

where

$$\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right] \;\equiv\; \mathbb{E}\left[\mathbb{E}_{\beta,\Lambda}\left[E\,\middle|\,M=\mu\right]\frac{f^{\Theta}_{\beta,\Lambda}(\mu)}{f^{\Theta}_{\beta}(\mu)}\right].$$

This definition is well defined by Result C.14. It is useful to show that it is strictly positive and bounded by one.

Result C.15. *For any $\beta \in (0,1)$ and any real μ, the expectation $\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right]$ is strictly positive and bounded above by one.*

Proof. Positivity is is a straightforward consequence of Assumption C.16 and Results C.10 and C.14. To see that $\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right] \leq 1$, note that

$$\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right] \;=\; \int \mathbb{E}_{\beta,\Lambda}\left[E\,\middle|\,M=\mu\right]\frac{f^{\Theta}_{\beta,\Lambda}(\mu)}{\int f^{\Theta}_{\beta}(\mu)\, dG^{\Lambda}(\lambda)}\, dG^{\Lambda}(\lambda).$$

The fraction $f^{\Theta}_{\beta,\Lambda}(\mu)\,/\int f^{\Theta}_{\beta}(\mu)\, dG^{\Lambda}(\lambda)$ is a density with respect to $G^{\Lambda}(\lambda)$, so $\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right]$ is an average of terms that are between 0 and 1. Hence, $\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right] \leq 1$. □

The expectation $\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right]$ is best thought of as a Λ-averaged and re-weighted version of $\mathbb{E}_{\beta,\lambda}\left[E\,\middle|\,M=\mu\right]$. The re-weighting factor, $f^{\Theta}_{\beta,\lambda}(\mu)\,/\,f^{\Theta}_{\beta}(\mu)$ ensures that if Λ causes a certain threshold to be more likely, then that threshold's expected update strength is more heavily weighted.

Putting our two pieces together, note that

$$\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right] f^{\Theta}_{\beta}(\mu) \;=\; \mathbb{E}\left[\mathbb{E}_{\beta,\Lambda}\left[E\,\middle|\,M=\mu\right] f^{\Theta}_{\beta,\Lambda}(\mu)\right],$$

that is, $\mathbb{E}_\beta\left[E \,\middle|\, M=\mu\right] f^\Theta_\beta(\mu)$ is the expectation over Λ of $\mathbb{E}_{\beta,\Lambda}\left[E \,\middle|\, M=\mu\right] f^\Theta_{\beta,\Lambda}(\mu)$. One might then think that $\mathbb{E}_\beta\left[E \,\middle|\, M=\mu\right] f^\Theta_\beta(\mu)$ inherits the monotone-likelihood-ratio property of Assumption C.17. This is not the case; we must assume it.

Assumption C.20. *For any two real numbers μ' and μ, and any two baselines, β' and β, on $(0,1)$, if $\mu' > \mu$ and $\beta' > \beta$, then the inequality*

$$\left\{\mathbb{E}_\beta\left[E \,\middle|\, M=\mu'\right] f^\Theta_\beta(\mu')\right\}\left\{\mathbb{E}_{\beta'}\left[E \,\middle|\, M=\mu\right] f^\Theta_{\beta'}(\mu)\right\} > \left\{\mathbb{E}_{\beta'}\left[E \,\middle|\, M=\mu'\right] f^\Theta_{\beta'}(\mu')\right\}\left\{\mathbb{E}_\beta\left[E \,\middle|\, M=\mu\right] f^\Theta_\beta(\mu)\right\}.$$

is satisfied.

The reason we must make Assumption C.20 is simple: Featherstone (2024) shows that the monotone-likelihood property doesn't necessarily aggregate for *arbitrary* distributions over Λ. That paper also provides a simple mixing condition that ensures it does; Assumption C.20 can be interpreted as assuming that mixing condition.

Now that we have discussed averaging out the literature-level noise, we are prepared to discuss how the expected treatment effect changes as a function of the baseline.

Expected treatment effect as baseline approaches 0 and 1

We begin by characterizing the convergence of F^Θ_β for baselines approaching 0 and 1.

Result C.16. *For any real θ, the following limits hold:*

$$\lim_{\beta\to 0} F^\Theta_\beta(\theta) \;=\; \lim_{\beta\to 1}\left(1 - F^\Theta_\beta(\theta)\right) \;=\; 0.$$

In other words, as β approaches 0 or 1, the measure represented by F^Θ_β converges vaguely to the zero measure.[9a]

Proof. We'll start with the proof for a baseline of 0; the proof for a baseline of 1 is completely symmetric. Consider a sequence (β^k) whose limit is zero. Clearly, $\liminf F^\Theta_{\beta^k}(\theta) \geq 0$, since distribution functions are non-negative.

9a. Vague convergence is a generalization of weak convergence that allows convergence to *sub*-probability measures that assign a less-than-one measure to the entire outcome space. The canonical example considers a sequence of random variables whose distributions are $\Phi(n-x)$. This sequence of distributions converges pointwise to zero as $n\to\infty$, but zero is not a distribution. So, the sequence fails to converge *weakly* to anything, but it does converge *vaguely* to the zero measure. See Chung (2001).

It is also true that $\limsup F^{\Theta}_{\beta^k}(\theta) \leq 0$. To see this, first note that, since distribution functions are weakly increasing, for any $K > 0$, there exists a k such that $F^{\Theta}_{\beta^k}(x) \geq \limsup F^{\Theta}_{\beta^k}(\theta)$ for any $x \geq \theta$. This implies we would have

$$\limsup \int_{-\infty}^{\infty} F^{\Theta}_{\beta^k}(\mu)\, d\Phi(\mu) \geq \left(\limsup F^{\Theta}_{\beta^k}(\theta)\right)\left(1 - \Phi(\theta)\right).$$

If $\limsup F^{\Theta}_{\beta^k}(\theta)$ were strictly positive, this would contradict the assumption that $\beta^k \to 0$. Hence, $\limsup F^{\Theta}_{\beta^k}(\theta) = 0$, which means that $F^{\Theta}_{\beta^k}(\theta) \to 0$. And since there was nothing special about θ, this must hold for for all θ on the real line. In other words, $F^{\Theta}_{\beta^k}$ converges *vaguely* to the zero measure chung2001course. □

Given this result, we can characterize the limiting behavior of the expected treatment effect as the baseline approaches 0 or 1.

Proposition C.4. *The limit of* $\mathbb{E}\left[\tau(N, B, \Lambda) \,\middle|\, B = \beta\right]$ *as* β *approaches either* 0 *or* 1 *is zero.*

Proof. Looking at Equation C.33, we can use the triangle inequality to write the bound

$$\left|\mathbb{E}\left[\tau(N, B, \Lambda) \,\middle|\, B = \beta\right]\right| \leq \int_{-\infty}^{\infty} \left(\bar{\nu} + |\mu|\right) \varphi(\mu)\, dF^{\Theta}_{\beta}(\mu).$$

where $\bar{\nu}$ is the bound from Assumption C.14. We have removed $\mathbb{E}_{\beta}\left[E \,\middle|\, M = \mu\right]$ as, by Result C.14, it is positive and bounded by 1.

Clearly, the integrand approaches zero as μ approaches $\pm\infty$. Then, using Result C.16 and the vague-convergence equivalent of the Portmanteau Theorem (Chung 2001, Theorem 4.4.1), $\left|\mathbb{E}\left[\tau(N, B, \Lambda) \,\middle|\, B = \beta\right]\right|$ must approach zero as β approaches 0 or 1. Hence, $\mathbb{E}\left[\tau(N, B, \Lambda) \,\middle|\, B = \beta\right]$ does as well. □

The expected treatment effect takes both signs

For a given realization, (ν, β, λ), of (N, B, Λ), when the inequality in Equation C.30 is satisfied, the treatment effect is strictly positive; otherwise, it is not. Looking to Equations C.32 and C.33, it is clear that the analogous condition for the conditional-on-baseline expected treatment effect to be strictly positive is

$$\mathbb{E}\left[N \,\middle|\, B = \beta\right] \;>\; \frac{\int_{-\infty}^{\mu} \tilde{\mu}\, \mathbb{E}_{\beta}\left[E \,\middle|\, M = \tilde{\mu}\right] f^{\Theta}_{\beta}(\tilde{\mu})\, \varphi(\tilde{\mu})\, d\tilde{\mu}}{\int_{-\infty}^{\infty} \mathbb{E}_{\beta}\left[E \,\middle|\, M = \tilde{\mu}\right] f^{\Theta}_{\beta}(\tilde{\mu})\, \varphi(\tilde{\mu})\, d\tilde{\mu}}. \tag{C.34}$$

The right-hand side of this inequality can be written as the first moment of a random variable whose density with respect to the measure F^{Θ}_{β} is

$$\psi_{\beta}(\mu) \equiv \frac{\mathbb{E}_{\beta}\left[E \,\middle|\, M = \mu\right] \varphi(\mu)}{\int_{-\infty}^{\infty} \mathbb{E}_{\beta}\left[E \,\middle|\, M = \tilde{\mu}\right] \varphi(\tilde{\mu})\, dF^{\Theta}_{\beta}(\tilde{\mu})};$$

its distribution is hence $\Psi_{\beta}(\mu) \equiv \int_{-\infty}^{\mu} \psi_{\beta}(\tilde{\mu})\, dF^{\Theta}_{\beta}(\tilde{\mu})$. This random variable is some sort of expectation-weighted prior, so we refer to it as $\mathbb{E}_{\beta}\left[M_E\right] \equiv \int_{-\infty}^{\infty} \mu\, \psi_{\beta}(\tilde{\mu})\, dF^{\Theta}_{\beta}(\tilde{\mu})$.

Clearly, the measure Ψ_β is absolutely continuous with respect to the measure F^Θ_β. We will need this to hold uniformly across all baselines to ensure the measure Ψ_β doesn't develop any atoms as β approaches zero or one. Formally,

Assumption C.21 (Uniform absolute continuity)**.** *For any $\iota > 0$, there exists an $\alpha > 0$ such that, for any set, A, and baseline, β, if $\int_A dF^\Theta_\beta < \alpha$, then, $\int_A \psi_\beta(\mu)\, dF^\Theta_\beta < \iota$.*

This assumption immediately implies that Ψ_β and F^Θ_β have the same behavior in the limit where the baseline approaches zero or one.

Result C.17. *For any real μ, the following limits hold:*

$$\lim_{\beta \to 0} \Psi_\beta(\mu) \;=\; \lim_{\beta \to 1} \left(1 - \Psi_\beta(\mu)\right) \;=\; 0.$$

Proof. Consider the set $A = (-\infty, \mu]$. By Assumption C.21, for any $\iota > 0$, there exists an $\alpha > 0$ such that when $F^\Theta_\beta(\mu) < \alpha$, $\Psi_\beta(\mu) < \iota$. But by Result C.16, for sufficiently small β, $F^\Theta_\beta(\mu) < \alpha$. Hence, for sufficiently small β, $\Psi_\beta(\mu) < \iota$. And since ι can be made arbitrarily small, we have established the first desired limit. The second can be proven in a similar manner, starting with the set $A = (\mu, \infty)$. □

By Assumption C.20, we see that $\Psi_\beta(\mu)$ is increasing in β, which will allow us to use the monotone-convergence theorem in what follows.

Result C.18. *For any real μ, the distribution function $\Psi_\beta(\mu)$ is strictly increasing in β.*

Proof. Assumption C.20 tells us that, for any two real numbers μ'' and μ, and any two baselines, β' and β, on $(0,1)$, if $\mu'' > \mu$ and $\beta' > \beta$, then the inequality

$$\begin{aligned}\left\{\mathbb{E}_\beta\left[E \,\middle|\, M = \mu''\right] f^\Theta_\beta(\mu'')\right\} \left\{\mathbb{E}_{\beta'}\left[E \,\middle|\, M = \mu\right] f^\Theta_{\beta'}(\mu)\right\} \\ > \left\{\mathbb{E}_{\beta'}\left[E \,\middle|\, M = \mu''\right] f^\Theta_{\beta'}(\mu'')\right\} \left\{\mathbb{E}_\beta\left[E \,\middle|\, M = \mu\right] f^\Theta_\beta(\mu)\right\}.\end{aligned}$$

is satisfied. If we integrate over all (μ, μ'') pairs that satisfy $\mu'' > \mu' > \mu$, we get

$$\begin{aligned}\int_{\mu'}^{\infty} \mathbb{E}_\beta\left[E \,\middle|\, M = \tilde{\mu}\right] f^\Theta_\beta(\tilde{\mu})\, d\Phi(\tilde{\mu}) \int_{-\infty}^{\mu'} \mathbb{E}_{\beta'}\left[E \,\middle|\, M = \tilde{\mu}\right] f^\Theta_{\beta'}(\tilde{\mu})\, d\Phi(\tilde{\mu}) \\ > \int_{\mu'}^{\infty} \mathbb{E}_{\beta'}\left[E \,\middle|\, M = \tilde{\mu}\right] f^\Theta_{\beta'}(\tilde{\mu})\, d\Phi(\tilde{\mu}) \int_{-\infty}^{\mu'} \mathbb{E}_\beta\left[E \,\middle|\, M = \tilde{\mu}\right] f^\Theta_\beta(\tilde{\mu})\, d\Phi(\tilde{\mu}).\end{aligned}$$

If we then divide by

$$\int_{-\infty}^{\infty} \mathbb{E}_{\beta'}\left[E \,\middle|\, M = \tilde{\mu}\right] \varphi(\tilde{\mu})\, dF^\Theta_{\beta'}(\tilde{\mu}) \int_{-\infty}^{\infty} \mathbb{E}_\beta\left[E \,\middle|\, M = \tilde{\mu}\right] \varphi(\tilde{\mu})\, dF^\Theta_\beta(\tilde{\mu}),$$

this becomes

$$\left(1 - \Psi_\beta(\mu')\right) \Psi_{\beta'}(\mu') \;>\; \left(1 - \Psi_{\beta'}(\mu')\right) \Psi_\beta(\mu'),$$

which simplifies to

$$\Psi_{\beta'}(\mu') > \Psi_\beta(\mu'),$$

as we set out to show. □

Finally, we are ready to state the main result from this section.

Result C.19. *The limits*

$$\lim_{\beta \to 0} \mathbb{E}_\beta \left[M_E \right] = +\infty \quad \textit{and}$$
$$\lim_{\beta \to 1} \mathbb{E}_\beta \left[M_E \right] = -\infty$$

are satisfied.

Proof. We will show just the first limit; the second follows from similar reasoning. Recall the integrated-tail-probability expectation formula (Lo, 2018), which states that

$$\mathbb{E}_\beta \left[M_E \right] = \int_0^\infty \left(1 - \Psi_\beta(\mu) \right) d\mu - \int_{-\infty}^0 \Psi_\beta(\mu) d\mu.$$

The first integral on the right-hand side is the expectation of the positive part of the random variable (i.e., $\mathbb{E}_\beta \left[M_E^+ \right] \equiv \mathbb{E}_\beta \left[\max\{0, M_E\} \right]$), while the second is the expectation of the negative part (i.e., $\mathbb{E}_\beta \left[M_E^- \right] \equiv \left| \mathbb{E}_\beta \left[\min\{0, M_E\} \right] \right|$).

The limit of $\mathbb{E}_\beta \left[M_E^+ \right]$ is ∞. To see this, consider a sequence of β_n that converge monotonically to zero. Then, $1 - \Psi_{\beta_n}(\mu)$ is a monotonically increasing sequence of non-negative functions that converges to one everywhere. Then, the monotone convergence theorem states that

$$\lim_{n \to \infty} \int_0^\infty \left(1 - \Psi_{\beta_n}(\mu) \right) d\mu = \int_0^\infty d\mu = \infty.$$

This states that for each $M > 0$, there exists $N > 0$ such that, when $n > N$, $\int_0^\infty \left(1 - \Psi_{\beta_n}(\mu) \right) d\mu > M$. But since this holds for *any* starting sequence, (β_n), that converges monotonically to 0, we have that when $\beta < \beta_N$, $\int_0^\infty \left(1 - \Psi_{\beta_n}(\mu) \right) d\mu > M$, which establishes the desired limit.

The limit of $\mathbb{E}_\beta \left[M_E^- \right]$ as β approaches zero is 0. To see this, using the same sort of monotonic sequence we used in the previous paragraph, note that $\Psi_{\beta_1}(\mu) - \Psi_{\beta_n}(\mu)$ is a monotonically increasing sequence of non-negative functions that converges to $\Psi_{\beta_1}(\mu)$. Then, we can use the same monotone-convergence-theorem logic to state

$$\lim_{n \to \infty} \int_{-\infty}^0 \left(\Psi_{\beta_1}(\mu) - \Psi_{\beta_n}(\mu) \right) d\mu = \int_{-\infty}^0 \Psi_{\beta_1}(\mu) d\mu.$$

Using the linearity of the integral, this becomes

$$\mathbb{E}_{\beta_1} \left[M_E \right] - \lim_{n \to \infty} \int_{-\infty}^0 \Psi_{\beta_n}(\mu) d\mu = \mathbb{E}_{\beta_1} \left[M_E \right],$$
$$\lim_{n \to \infty} \int_{-\infty}^0 \Psi_{\beta_n}(\mu) d\mu = 0.$$

Plugging the results from the past two paragraphs into the integrated-tail-probability expectation formula, we have shown that $\mathbb{E}_\beta \left[M_E \right]$ approaches ∞ as β approaches zero. □

This, in turn, allows us to know that the treatment effect takes on both positive and negative values.

Proposition C.5. *There is some $\underline{\beta} > 0$ such that, when $\beta < \underline{\beta}$, $\mathbb{E}\left[\tau(N,B,\Lambda)\,\middle|\,B=\beta\right] < 0$. Similarly, there is some $\bar{\beta}$ such that, when $\beta > \bar{\beta}$, $\mathbb{E}\left[\tau(N,B,\Lambda)\,\middle|\,B=\beta\right] > 0$.*

Proof. Assumption C.14 states that $\mathbb{E}\left[N\,\middle|\,B=\beta\right]$ is bounded above and below. Result C.19 states that as $\beta \to 0$, $\mathbb{E}_\beta\left[M_E\right]$ grows without bound. Hence, there exists some $\underline{\beta}$ such that, for all $\beta < \underline{\beta}$, the inequality in Equation C.34 will eventually be violated, leading $\mathbb{E}\left[\tau(N,B,\Lambda)\,\middle|\,B=\beta\right]$ to be negative. A similar logic shows that $\mathbb{E}\left[\tau(N,B,\Lambda)\,\middle|\,B=\beta\right]$ is positive for all baselines that exceed some $\bar{\beta}$. □

Single crossing of the expected treatment effect

We now turn to show that the conditional-on-baseline expected treatment effect is single crossing from below.

Proposition C.6. *Take two baselines on $(0,1)$, β' and β, where $\beta' > \beta$. When $\mathbb{E}\left[\tau(N,B,\Lambda)\,\middle|\,B=\beta\right] \geq 0$, it must be that $\mathbb{E}\left[\tau(N,B,\Lambda)\,\middle|\,B=\beta'\right] > 0$.*

Proof. Define

$$\tau^+(\beta',\beta) \equiv \int_{-\infty}^{\mathbb{E}\left[N\,\middle|\,B=\beta\right]} \mathbb{E}_{\beta'}\left[E\,\middle|\,M=\mu\right]\, f^{\Theta}_{\beta'}(\mu)\, \left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu\right) d\Phi(\mu),$$

$$\tau^-(\beta',\beta) \equiv \int_{\mathbb{E}\left[N\,\middle|\,B=\beta\right]}^{\infty} \mathbb{E}_{\beta'}\left[E\,\middle|\,M=\mu\right]\, f^{\Theta}_{\beta'}(\mu)\, \left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu\right) d\Phi(\mu), \quad \text{and}$$

$$\tau(\beta',\beta) \equiv \int_{-\infty}^{\infty} \mathbb{E}_{\beta'}\left[E\,\middle|\,M=\mu\right]\, f^{\Theta}_{\beta'}(\mu)\, \left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu\right) d\Phi(\mu),$$

so that $\tau(\beta',\beta) = \tau^+(\beta',\beta) + \tau^-(\beta',\beta)$. Then, $\tau(\beta,\beta)$ is the expected treatment effect conditional on the baseline being β (i.e., $\mathbb{E}\left[\tau(N,B,\Lambda)\,\middle|\,B=\beta\right]$), and $\tau^+(\beta,\beta)$ and $\tau^-(\beta,\beta)$ are the respective contributions to the treatment effect from those who, in expectation, are nudged into and out of take-up.

For all (μ,μ') pairs that satisfy $\mu < \mathbb{E}\left[N\,\middle|\,B=\beta\right] < \mu'$, Assumption C.20 holds. If we then multiply it by $\left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu\right)\,\left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu'\right)$, the inequality flips, yielding

$$\begin{aligned}
&\left\{\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu'\right] f^{\Theta}_{\beta}(\mu')\right\}\,\left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu'\right)\\
&\quad\left\{\mathbb{E}_{\beta'}\left[E\,\middle|\,M=\mu\right] f^{\Theta}_{\beta'}(\mu)\right\}\,\left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu\right)\\
&\qquad< \left\{\mathbb{E}_{\beta'}\left[E\,\middle|\,M=\mu'\right] f^{\Theta}_{\beta'}(\mu')\right\}\,\left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu'\right)\\
&\qquad\quad\left\{\mathbb{E}_{\beta}\left[E\,\middle|\,M=\mu\right] f^{\Theta}_{\beta}(\mu)\right\}\,\left(\mathbb{E}\left[N\,\middle|\,B=\beta\right]-\mu\right).
\end{aligned}$$

If we then integrate with respect to the measure Φ over all (μ,μ') pairs that satisfy $\mu < \mathbb{E}\left[\tau(N,B,\Lambda)\,\middle|\,B=\beta\right] < \mu'$, we get

$$\tau^-(\beta,\beta)\,\tau^+(\beta',\beta) \;<\; \tau^-(\beta',\beta)\,\tau^+(\beta,\beta)$$

Since the τ^- terms are negative, we can also write

$$|\tau^-(\beta,\beta)|\ \tau^+(\beta',\beta) \ > \ |\tau^-(\beta',\beta)|\ \tau^+(\beta,\beta),$$

Now, an immediate consequence of Results C.14 and C.15 is that $\tau^+(\beta',\beta)$ and $\tau^-(\beta',\beta)$ are both non-zero. When $\beta' = \beta$, this says that there is some mass of the population that the expected signal, $\mathbb{E}\left[N \middle| B = \beta\right]$, would nudge into take-up and some mass that it would nudge out of take up.

Given the positivity of $\tau^-(\beta',\beta)$ and $\tau^-(\beta,\beta)$, we can transform the previous inequality to

$$\frac{\tau^+(\beta',\beta)}{|\tau^-(\beta',\beta)|} \ > \ \frac{\tau^+(\beta,\beta)}{|\tau^-(\beta,\beta)|}.$$

By assumption, $\mathbb{E}\left[\tau(N,B,\Lambda) \middle| B = \beta\right] \geq 0$, which is equivalent to $\tau^+(\beta,\beta) - |\tau^-(\beta,\beta)| \geq 0$, which is equivalent to $\tau^+(\beta,\beta)/|\tau^-(\beta,\beta)| \geq 1$. Hence, $\tau^+(\beta',\beta)/|\tau^-(\beta',\beta)| > 1$, which tells us that $\tau^+(\beta',\beta) - |\tau^-(\beta',\beta)| > 0$, which is equivalent to $\tau(\beta',\beta) > 0$.

Assumption C.18 immediately implies that $\mathbb{E}\left[N \middle| B = \beta\right]$ is weakly increasing in β. So, $\tau(\beta',\beta)$ is weakly increasing in its second argument, which tells us that $\tau(\beta',\beta') \geq \tau(\beta',\beta)$. Hence, if $\tau(\beta',\beta)$ is strictly positive, then so is $\tau(\beta',\beta')$. But, $\tau(\beta',\beta') = \mathbb{E}\left[\tau(N,B,\Lambda) \middle| B = \beta'\right]$, so we have shown that $\mathbb{E}\left[\tau(N,B,\Lambda) \middle| B = \beta'\right] > 0$, establishing the desired result. □

The shape of the expected treatment effect curve

Results C.4, C.5, and C.6 establish the shape of the conditional-on-baseline expected treatment effect.

Theorem C.3. *$\mathbb{E}\left[\tau(N,B,\Lambda) \middle| B = \beta\right]$ approaches zero as β approaches zero or one. Further, there is some baseline, $\beta_0 \in (0,1)$, such that $\mathbb{E}\left[\tau(N,B,\Lambda) \middle| B = \beta\right] \leq 0$ when $\beta \leq \beta_0$ and $\mathbb{E}\left[\tau(N,B,\Lambda) \middle| B = \beta\right] > 0$ when $\beta > \beta_0$.*

Pictorially then, we expect something like part (h) of Figure 1 in the main text.