The non-linear effect of descriptive social norms*

Jesper Akesson, Rena Conti, Robert W. Hahn,
Robert D. Metcalfe and Itzhak Rasooly†

August 5, 2021

Abstract

In this paper, we conduct a field experiment in which different participants are sent different signals about the share who sign up as advocates for a non-profit organization. Participants are then given the opportunity to themselves sign up as advocates. We find that while sending higher signals generally increases the share who sign up, this relationship reverses at low signal levels. We discuss what might explain this result as well as its implications for the effective design of social norm interventions.

Keywords: social norms, field experiment, beliefs
JEL codes: D90, C93
Word count: 5689

*We would especially like to thank Elisa Weiss, Maria Sae-Hau, Gabriela Gracia, and the team at the Leukemia & Lymphoma Society for helping us design and conduct this experiment. We would also like to thank Filippo Muzi-Falconi and Ondrej Kacha for excellent research assistance, as well as Simge Andi, Nikhil Kalyanpur, Luke Milsom, JingKai Ong, Chris Roth, Toby Shevlane and Jasmine Theilgaard for helpful comments and suggestions. Boston University IRB: A231602CD; AEA RCT registry: AEARCTR-0007938.

†Akesson, The Behaviouralist; Conti, Boston University; Hahn, University of Oxford and Technology Policy Institute; Metcalfe, University of Southern California and NBER; Rasooly, University of Oxford.

1 Introduction

Descriptive social norm interventions are attempts to influence behavior by changing individual beliefs about the prevalence of an activity. They have been used in several areas, including tax compliance (Hallsworth et al., 2017; Perez-Truglia and Troiano, 2018), online reviews (Chen et al., 2010), alcohol consumption (Borsari and Carey, 2003), voting (Gerber and Rogers, 2009), labor force participation (Bursztyn et al., 2020), and resource conservation (Allcott, 2011; Ferraro and Price, 2013; Brent et al., 2015). The body of work that has emerged from these interventions reveals that they can exert a powerful influence on behavior.

Despite the large number of papers on this topic, at least two prominent gaps remain. First, there are very few papers that examine the relative effects of sending different (exogenous) signals about the prevalence of a behavior (e.g., that 80% vs 90% pay their taxes on time). Moreover, those papers that do (e.g., Frey and Meier (2004)) only compare the relative effects of sending two different signal levels.[1] This makes it difficult to understand how the effects of these interventions depend on the exact signal they convey and when such interventions should be used.

Second, there appear to be no descriptive social norm experiments that examine the effect of suggesting that no one else is doing an activity. Some scholars argue that descriptive norm interventions should not be used in low-compliance environments, as they might simply reinforce undesirable behaviors (e.g., Bicchieri and Dimant (2019)). However, such assertions have not been rigorously tested, and there are theoretical reasons why it may be beneficial to convey information about low levels of compliance. For example, it is possible that individuals believe that the impact of their action is lower the greater the number of others who are also engaging in the activity. Alternately, decreasing beliefs regarding the share of individuals that engage in an altruistic activity might increase the perceived 'self-image' benefits from engaging in that activity (Bénabou and Tirole, 2006; Falk, 2020).

In this paper, we address both of the gaps in the literature discussed above. To do this, we conduct a natural field experiment ($n = 5,360$) in partnership with the Leukemia &

[1] See Linek and Traxler (2021)'s Trials 3-4 and Heldt (2005) for other experiments that study the relative impact of two different signal values, albeit in contexts that are different from ours. Two more distantly related studies are Perez-Truglia and Troiano (2018) and Perez-Truglia and Cruces (2017). In the former, the authors randomly vary the signals that participants receive regarding the amount that subjects' neighbors owe. In the latter, the authors randomly vary signals about how much participants' neighbors have donated to a political campaign. Neither of these papers vary information about the share of individuals to engage in a particular activity (as in our study).

Lymphoma Society (LLS). Members of the society are randomly placed into one of six treatment groups or a control group. Each of the treatment groups is given a different non-deceptive signal about the share of members who had signed up as advocates for the organization (ranging from 0% to 100% in 20% increments).[2] Subjects are then offered the opportunity to sign up as advocates, which LLS views as a key to achieving its goals. This design allows us to study precisely how sign-up rates vary with the information conveyed by the intervention. By collecting data on individual beliefs both before and after the intervention, we are also able to explore the mechanisms through which the observed effects occur.

Our analysis yields three main findings. First, we find that signals about the share of individuals who sign up exerts a powerful influence on behavior in this context. For example, while only 22% of those in the control group (who did not receive any information) sign up, this share rises to 30% in the group that received a signal indicating that everybody signed up ($p = 0.02$). Our study thus adds to the growing body of work demonstrating the power of descriptive social norms, and contributes to the literature on how to encourage political participation (see, e.g., Gerber and Rogers (2009)).

Second, we show that the relationship between signal and behavior is convex. While increasing the signal generally increases sign-up rates, sign-up rates actually fall when the signal increases from 0% to 20%. To our knowledge, this is the first experimental evidence suggesting that descriptive social norms can increase compliance by revising beliefs about prevalence downward.

We then ask what generates the convex relationship that we observe. This yields our third finding: the convexity does not stem from a convex relationship between signals and beliefs, but rather from a convex relationship between beliefs and behavior. In other words, there appear to be 'increasing returns' to revising beliefs about prevalence upward over certain domains. We conduct this analysis using instrumental variables, with our treatments instrumenting for beliefs (see, e.g., Haaland et al. (2020) for other examples of this approach).

Furthermore, we explore why sending a very low signal increases compliance in our setting. While we are unable to definitively identify the cause, we are able to reject the main 'rational' hypothesis: that people believe that the impact of their advocacy is larger if very few others sign up as advocates. We propose that this effect is instead driven by either self-image concerns or feelings of pity and guilt.

[2] LLS advocates work with the society to influence policy on federal and state levels. Advocates may be asked to attend events or write letters to their members of Congress. See https://www.lls.org/policy-advocacy for more information.

Finally, we study the extent to which our descriptive social norm interventions work through revising beliefs about what others are doing (which we mainly capture using elicited point estimates). While it is natural to think that this is the primary channel through which they operate, other channels are possible. For example, it might be that instead of revising individual beliefs (i.e., point estimates), the interventions simply make certain beliefs more salient. Indeed, our analysis suggests that other channels may be operative: even controlling for beliefs, we find that some signals still have significant effects on behavior. However, a simple calculation shows that changes in beliefs account for the majority of the observed effects, suggesting that belief updating is the main mechanism underlying our results.

The study that most closely relates to our research appears to be Linek and Traxler (2021). In a series of trials, the authors send different signals about the share who donate to Wikipedia and examine how these signals impact donation rates. In contrast to our study, Linek and Traxler (2021) do not compare the effect of these signals relative to a control group that receives no information. In addition, their social information interventions do not seem to have shifted individual beliefs about donations—which may explain why they, unlike us, obtain null effects of varying the signal.

The remainder of this paper is structured as follows. Section 2 contains the experimental design and Section 3 presents the effect of the interventions on the share that sign up as advocates. This effect is then decomposed into the effect of the interventions on beliefs (Section 4), and the effects of beliefs on sign-ups (Section 5). Section 6 explores why revising beliefs downward may increase sign-ups, and Section 7 considers the extent to which the effects are purely belief driven. Finally, Section 8 reviews our main findings and discusses policy implications.

2 Experimental design

The experiment was conducted in December 2020 in collaboration with the Leukemia & Lymphoma Society (LLS). At that time, LLS was conducting a survey about how COVID-19 had affected the lives of blood cancer patients, survivors, and caregivers in the US. We embedded our experiment within the survey that LLS emailed to its constituents.

In total, 5,360 individuals took part in the experiment, 63% of whom identified as female and 50% of whom reported holding a Bachelor's degree or higher. The average age of respondents was 64. The survey was administered using Qualtrics. Qualtrics was also used to conduct the randomization.

The experiment took place toward the end of the survey, and began by asking individuals if they were familiar with LLS advocacy work (around 69% stated that they were at least slightly familiar with these efforts). Participants were then provided with the following information about LLS 'advocates':

> *Members of the LLS Action Team play a critical role in our efforts to influence health policy at the federal and state levels. Advocates all over the country are sharing their stories and taking action on public policy issues that matter to blood cancer patients and survivors—from improving access to treatment to promoting meaningful insurance coverage for patients to accelerating new cures.*
>
> *LLS sends advocates simple ways to communicate with elected officials via email, phone calls and other actions. Their voice could be the one that sways a decision-maker to do the right thing.*

Participants were subsequently told that we had asked past respondents whether they would like to sign up to the LLS Action Team. They were then asked the following question:

> *Roughly, what percentage of past respondents to this survey do you think signed up and became advocates? Please provide us with your best guess (It's okay if you don't get it right!)*

Participants were able to enter any integer between 0 and 100.[3] We also elicited participants' degree of confidence about their response on a 5-point Likert scale (from 'very unconfident' to 'very confident').

Prior to conducting the experiment, we had used the results of a pilot survey ($n = 200$) to create six groups of participants, each with ten members.[4] The first group had been chosen so that none of the ten members had signed up; the second group had been chosen so that two of the ten members had signed up; and so on (so that, in every group, X of the 10 had signed up for some $X \in \{0, 2, ..., 10\}$). In the main survey, we then randomly matched each participant with one of these six groups (or instead allocated them to the control group who

[3]We did not pay individuals money for correctly answering this question since this would have seemed out of place within the context of an LLS COVID-19 survey, in part because many respondents were presumably answering the questions to aid cancer research. In fact, providing financial incentives may have crowded out their intrinsic motivation to answer accurately, thus worsening the quality of the data. Moreover, there is little evidence that incentivizing correct guesses makes individuals more likely to reveal their true beliefs, and even some evidence that it makes them less likely to do so (Haaland et al., 2020).

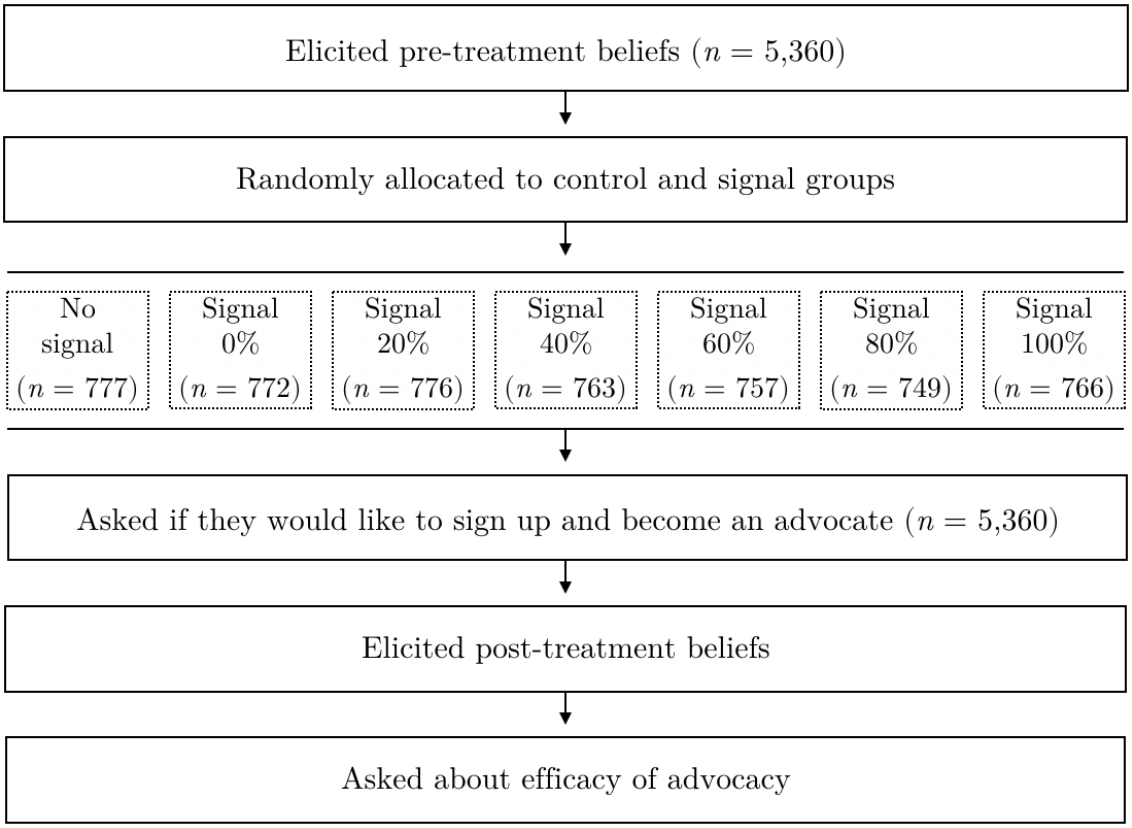
[4]The pilot experiment was identical to the main experiment except for the fact that individuals were not given an 'informational treatment'. Hence, participants in the pilot experiment completed the same survey as those in the control group of the main experiment. We do not use any data from the pilot in our analysis.

received no information).[5] Unless in the control group, each participant was then informed of the following:

> *We have matched you with 10 other members of the LLS community who have responded to this survey. X of these 10 decided to sign up and become advocates!*

As can be seen in Figure 1, around 760 participants were matched with each of the six groups that received a signal about the share who had signed up (or placed in the control group).[6] We then recorded our main outcome of interest, namely whether individuals signed up and became advocates. We did this by offering them the opportunity to sign up within the survey.

Figure 1: Experimental design



In contrast to several other descriptive social norm interventions that provide information about prevalence within the entire population (e.g., Andı and Akesson (2020) or Hallsworth et al. (2017)), we only provided information about a small sample of individuals. The

[5]Our strategy is thus in a similar spirit to Bérgolo et al. (2017), who also generate exogenous variation in perceptions without resorting to deception.

[6]See Table A1 in the Appendix for a balance table. The treatment arms are balanced on observables.

main drawback of this approach (relative to that taken by usual descriptive social norm interventions) is that it may limit the extent to which respondents update their beliefs about the share of LLS members who signed up. However, this approach allowed us to provide different, and yet still accurate information, to different respondents. In any case, the signals that we provided did update beliefs in the expected way (see Section 4) and in a way similar to other interventions that provide more representative information (see, e.g., Cullen and Perez-Truglia (2018)).

After having decided whether they would like to sign up or not, participants were again asked about their beliefs regarding the share that signed up. More specifically, we asked:

> *Based on what you now know, roughly what percentage of past respondents to this survey do you think signed up and became advocates?*

Respondents could enter any integer between 0 and 100, and we again asked about the extent to which they were confident in their response.

Finally, the experiment concluded by asking:

> *Do you agree with the following statement: I can help shape health policy and improve the lives of patients as a member of the LLS Action Team.*

Respondents provided their answer on a 5-point Likert scale (ranging from 'strongly agree' to 'strongly disagree'). We asked this particular question in an attempt to measure the extent to which the treatments operate by changing beliefs about the value of LLS advocacy.

3 Treatment effects on the share that sign up

In this section, we examine the effects of the six treatments on the share of participants that sign up and become advocates. In order to do this, we estimate the following model:

$$y_i = \beta_0 + \sum_{k=1}^{6} \beta_k T_{ki} + u_i \tag{1}$$

where y_i is a dummy variable indicating whether individual i signs up and the T_{ki} are dummy variables tracking the treatment (or signal) to which they are assigned.

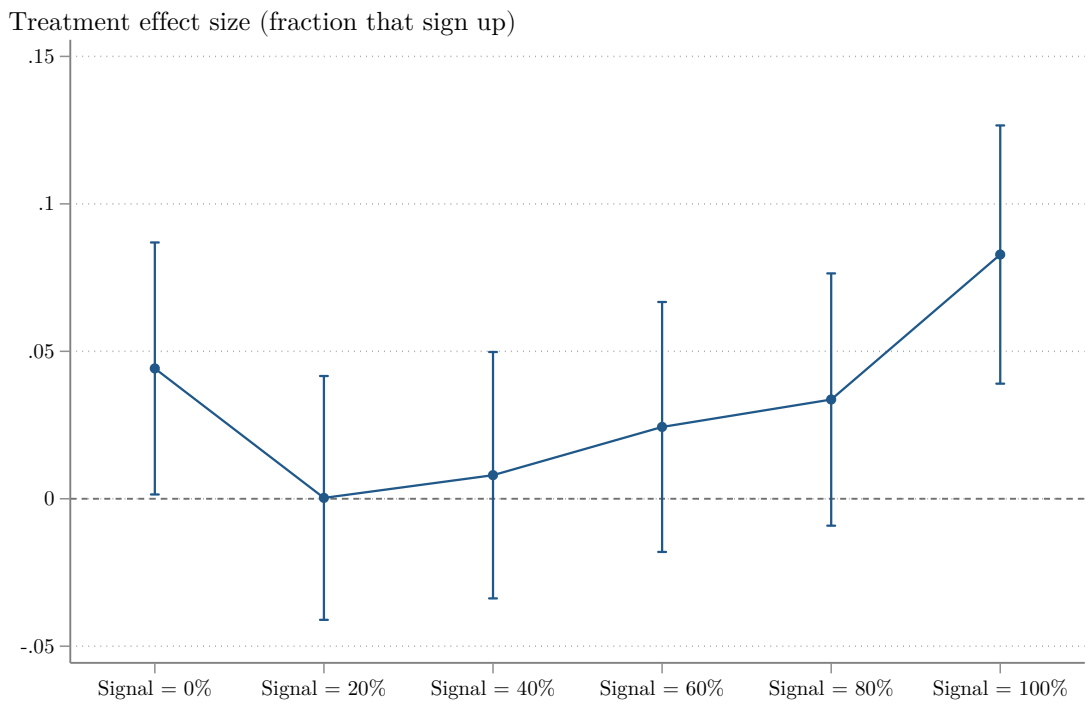
The results are presented in Figure 2 (and in Table A2 in the Appendix), with p-values indicating whether the proportion of participants who sign up differs significantly relative to those in control group. Being shown a signal of 100% has the greatest effect on sign-ups, bringing the share who sign up from 22% (in the control group) to 30% ($p < 0.001$). Perhaps

more surprisingly, the second-most effective signal is 0%, with an effect size of 4.4 percentage points relative to the control ($p = 0.04$). While the estimated coefficients for the 60% and 80% signals are positive and non-negligible, they are not statistically significant.

The effect of the signal on the share of participants that sign up does not seem to be linear. Instead, at first glance, it appears to be convex. To test this formally, note that linearity is equivalent to the restriction $\beta_{i+1} - \beta_i = \beta_{i+2} - \beta_{i+1}$ for all $i \in \{1, ..., 4\}$. Re-estimating the model under these constraints leads to a substantial worsening of fit; and indeed, an F-test strongly rejects the restrictions ($p < 0.01$). Since the relationship is clearly not concave, any departure from linearity should be interpreted as a convex relationship.

This finding raises the question of why the observed relationship is convex. In theory, this could be either due to a non-linear effect of the signals on participants' beliefs about the share that signed up; or instead a non-linear effect of beliefs on their behavior (or both). We attempt to answer this question in the following two sections.

Figure 2: The effects of different norm signals on the share that sign up



Notes: This figure presents regression coefficients for the model presented in Equation (1) with 95% confidence intervals. The regression was conducted using a Linear Probability Model. The y-axis represents the difference between the share that sign up in the control group and the respective treatment groups.

4 Treatment effects on beliefs

We begin by examining the effects of the treatments on posterior point estimates about prevalence (i.e., 'beliefs'). To this end, we estimate the model

$$b_i = \gamma_0 + \sum_{k=1}^{6} \gamma_k T_{ki} + u_i \tag{2}$$

where b_i is the posterior point estimate of individual i and the T_{ki} are dummy variables tracking the treatment to which they were assigned (as before).

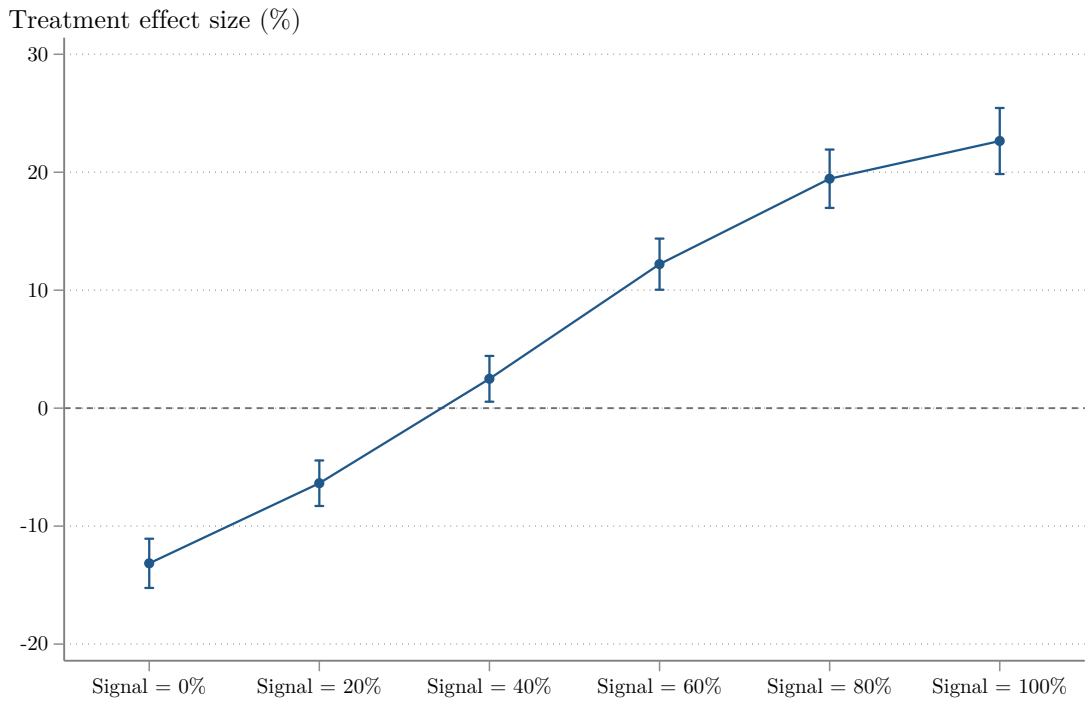
The results are displayed in Figure 3 (see also Table A4 which displays the exact numbers both with and without additional demographic control variables).[7] As can be seen, higher signals generally lead to higher beliefs. Indeed, we can reject the hypothesis that average beliefs are the same (with $p < 0.01$) for every 'neighbouring' treatment group except for the groups receiving the 80% and 100% signals; and can more weakly ($p < 0.05$) reject the hypothesis that they are the same for the 80% and 100% groups. However, while the relationship between signals and beliefs is positive, it certainly does *not* appear to be convex. On the contrary, it appears to be roughly linear over the 0% - 80% range only to flatten out at the end.[8] Thus, we do not yet have an explanation for the observed convex relationship between signals and behaviors.

Figure 3 reveals another interesting finding: for every group, the average posterior belief lies strictly between the average prior belief and the signal which that group was given. For instance, the average posterior belief in the 0% signal group was 18%, which lies between 0% (the signal) and 34% (the average prior for that group). While this certainly does not prove that individuals update their beliefs 'rationally', we note that it is consistent with simple models in which individuals update their beliefs in a Bayesian manner, placing some weight on their prior but also some weight on the signal they receive (see, e.g., Armantier et al. (2016) and Cullen and Perez-Truglia (2018)).

[7] In addition, Figures A1–A7 in the appendix display how the treatments alter the entire distribution of beliefs, not just their means.

[8] Indeed, we are not able to reject the hypothesis that $\gamma_{i+1} - \gamma_i = \gamma_{i+2} - \gamma_{i+1}$ for all $i \in \{1, 2, 3\}$ ($p = 0.995$).

Figure 3: The effects of different norm signals on beliefs about the share that sign up



Notes. This figure presents regression coefficients for the model presented in Equation (2) with 95% confidence intervals. The regression was conducted using OLS. The y-axis represents the difference between the post-treatment beliefs about share that sign up in the control group and the respective treatment groups.

5 Effects of beliefs on sign-ups

Having examined the effects of treatments on beliefs about prevalence, we now examine how these beliefs influence sign-up rates. In particular, we examine whether the effects of beliefs on behaviors explains the non-linearity documented in Section 3. In order to do so, we adopt an instrumental variables (IV) strategy where we use random treatment assignment as instruments for participants' beliefs (see Haaland et al. (2020), Cullen and Perez-Truglia (2018), Bottan and Perez-Truglia (2020), Roth and Wohlfart (2020) or Akesson et al. (2020) for examples of other studies that use this approach to estimate the effects of beliefs on behavior).

To be more concrete, we estimate the model

$$y_i = \theta_0 + \theta_1 b_i + \theta_2 b_i^2 + u_i \tag{3}$$

where y_i is a binary variable denoting whether individual i signs up and b_i is individual i's

belief (as before). We include the b_i^2 term to allow for possible non-linearity in the effect of beliefs on sign-ups.[9] We then instrument for the two explanatory variables using the treatments, so the 'first stage' equations take the form

$$b_i^k = \delta_0 + \sum_{k=1}^{5} \delta_k T_i + u_i \tag{4}$$

where the T_i are dummies indicating treatment assignment for $k = 1, 2$ (the zero signal forms the reference group). That is, we instrument for beliefs and beliefs squared using the 5 treatment assignment dummies. We omit the control group from this analysis to ensure that the exclusion restriction is satisfied (we discuss this issue in more detail later).

We begin by confirming that our instruments are informative (as one might expect in light of Figure 3). To do this, we conduct the standard eigenvalue test (Stock and Yogo, 2005) and can strongly reject the null hypothesis that the instruments are weak. We now proceed to consider whether our instruments are valid (i.e., whether the exclusion restriction holds).

It is unclear whether the exclusion restriction would hold (at least exactly) if we compare individuals in the treatment group to those in the control group. For example, providing individuals with information might not just influence their point estimates but may also change the 'salience' of these point estimates, implying that those who receive information (in the treatment groups) are differentially primed to those who do not (in the control). It is considerably more plausible, however, that the exclusion restriction holds when only considering our treatment groups. The *only* thing that varies across these groups is the signal presented, which means (for instance) that individuals in all groups should be equally 'primed'.

While the exclusion restriction is more plausible when considering only the treatment groups, it is possible that varying signals does not influence just individuals' point estimates but also the confidence which individuals place in these estimates. To test this, we estimate the effect of the treatments on individuals' confidence in their guess (measured on a five-point Likert scale). We do not find any significant effects and the estimated effects are close to zero (see Columns (1) and (2) of Table A3). This holds regardless of whether we treat confidence as a

[9] We also introduced higher order polynomials but these were not significant and did not substantially change our results—see Table A5.

continuous variable or collapse it into a binary variable (at any possible cut-off).[10] Moreover, our IV estimates (discussed later) do not change appreciably if confidence is controlled for. Thus, any possible 'confidence effect' is unlikely to be an issue. Since this is the only threat to the exclusion restriction which we could think of, we believe that it holds in this setting.

Figure 4 displays the results of estimating Equation (3) (see also Table A5). As can be seen, imposing a linear relationship between beliefs and sign-ups reveals a positive effect: increasing point estimates by 1 percentage point is predicted to increase sign-up rates by 0.1 percentage points ($p = 0.035$). However, this result disguises important non-linearities in the effect of beliefs on sign-ups. Indeed, we can clearly reject the hypothesis that the coefficient on the quadratic term is zero ($p < 0.001$). Moreover, Figure 4 suggests that the relationship between beliefs and sign-up rates is U-shaped. When taken in conjunction with the result from Section 4 (that the relationship between signals and beliefs is roughly linear), this finally provides an explanation for the pattern of treatment effects observed in Section 3.[11]

One might wonder whether the U-shape that we find is an artifact of the model that we specify in (3). However, including higher order polynomials does not substantially change the predicted relationship between beliefs and sign-ups (see Table A5). For example, including a cubic term results in near identical predicted sign-up rates for every value of beliefs. To further study the robustness of this finding, we also estimate the model

$$y_i = \lambda_0 + \sum_{k=0}^{4} \lambda_{k+1}(b_i - 20k)\mathbb{1}[b_i \geq 20k] \tag{5}$$

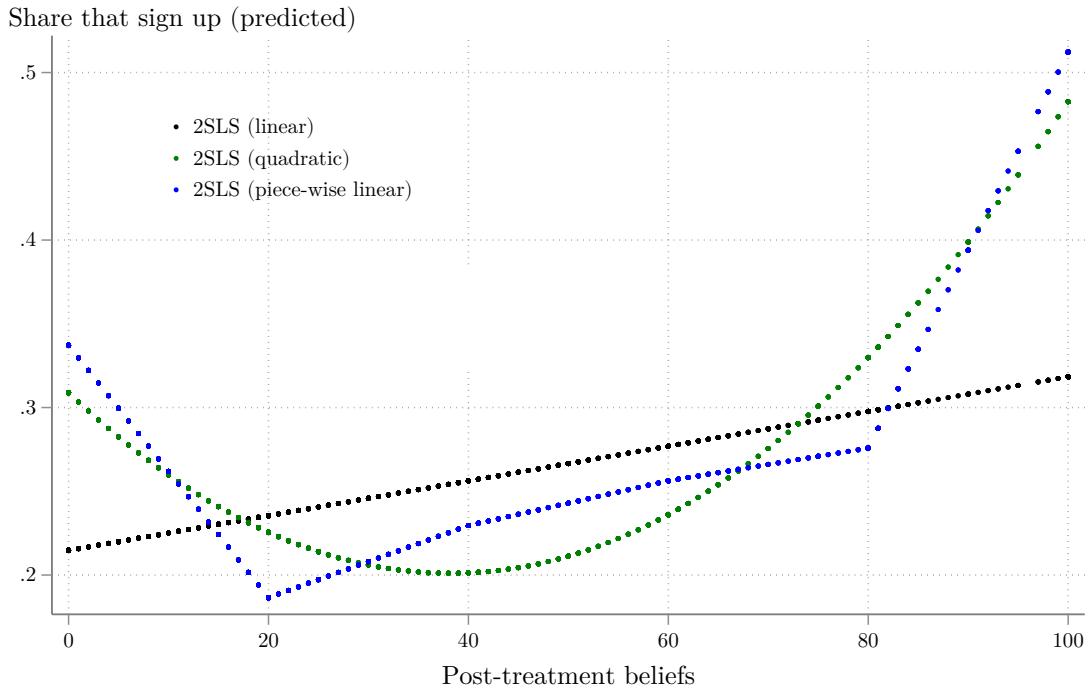
where $\mathbb{1}[b_i \geq 20k]$ is an indicator function that equals 1 if $b_i \geq 20k$ (and equals zero otherwise). In other words, we assume that the relationship between beliefs and sign-ups is 'piece-wise linear'; allowing the effect to vary as beliefs increase but ensuring that the relationship is continuous.[12] Figure 4 displays the results (see also Table A6 in the appendix).

[10] In contrast, we find that the treatments dramatically increase confidence relative to the control (see Columns (3) and (4) of Table A3); providing another reason to exclude the control group from this analysis. This finding emphasises the importance of including multiple treatment groups in information provision experiments; and suggests that findings that are based on information provision alone should be interpreted with caution.

[11] More formally, if a function f is linear in x, and another function g is convex in x, then the composed function $g \circ f$ is convex is x. So if beliefs are roughly linear in the signal value, and the sign-ups are convex in beliefs, sign-ups should be convex in the signal value—the exact phenomenon we observe.

[12] Since we have six treatment groups, we have five dummy variables indicating treatment assignment—which means that we can instrument for at most 5 explanatory variables. As a result, we allow the relationship between beliefs and sign-ups to vary over 5 intervals, estimating the most flexible possible model whose parameters can be identified using our data.

Figure 4: The effects of beliefs on behavior (predicted values)



Notes. In this figure, we present predictions from three instrumental variable regression models (which can be found in Tables A5 and A6). The dots represent the predicted sign-up rate for all individuals in the sample, given their beliefs about the share that sign up.

As can be seen, the estimated relationship is broadly similar to before (although the coefficient estimates are not significant). In particular, the model predicts that increasing beliefs decreases sign-ups when beliefs are low but increases sign-ups when beliefs are high. Thus, the results from this model are broadly consistent with those observed previously.[13]

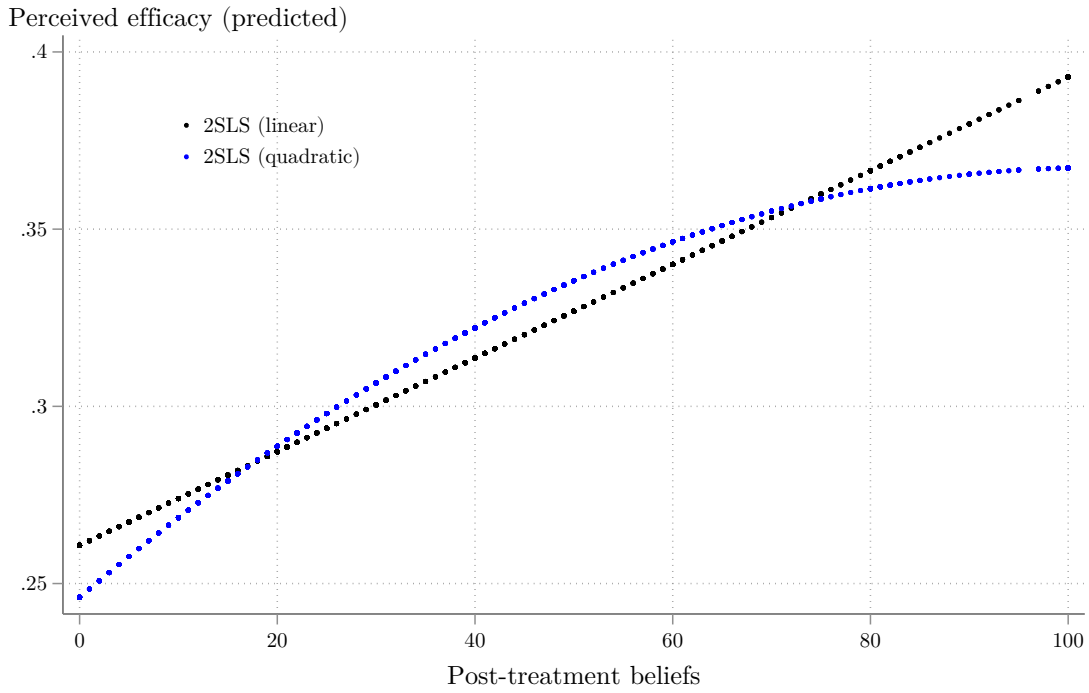
6 Why can revising beliefs downwards increase sign-ups?

The previous section documents that sign-ups can actually rise when beliefs about the share who sign up fall. Obviously, this is quite a surprise: for example, if individuals have a non-instrumental desire to conform or take higher sign-up rates as a signal of greater material benefits from signing up (Banerjee, 1992; Bikhchandani et al., 1992), then one would expect

[13] As a further robustness check, we estimate a similar model which includes a piece-wise linear relationship over four intervals and an additional dummy that equals one when beliefs are exactly zero. The purpose of this model is to check whether 'zero is special' and whether this might be driving our results. While it appears that zero may be qualitatively different from other beliefs (see Table A6), this specification does not generate a substantially different estimated relationship between beliefs and behavior.

higher beliefs about sign-up rates to lead to more sign-ups. In this section, we discuss what might explain this rather puzzling result (which to our knowledge has not been observed in previous descriptive social norm experiments).

Figure 5: The effects of beliefs on perceived efficacy (predicted values)



Notes. In this figure, we present predictions from two instrumental variable regression models (which can be found in Table A7). The dots represent the (predicted) share of respondents who believe that LLS advocacy is effective, given their level of beliefs about the share that sign up.

One potential explanation for the effect that we observe is the existence of (perceived) 'diminishing returns'. By this, we mean that individuals believe that the impact of their advocacy is lower the greater the number of others who are also advocates.[14] To test this, we examine our measure of participants' perceptions of the efficacy of being an advocate, estimating the exact same equations as in Section 5 but changing the dependent variable to the share that believe advocacy is effective. The results of these regressions are presented in Table A7 (and are plotted in Figure 5). As can be seen, perceived efficacy appears to be strictly increasing in beliefs about the share who sign up.[15] As a result, perceptions about diminishing returns do *not* seem to explain the results that we observe here: on the contrary,

[14]This might be, for example, because tasks are believed to be completed in order of decreasing importance and the value of each tasks is independent of how many other tasks are completed (see Lerner (1944) for a discussion of this point in the context of utility theory).

[15]This is precisely what one would expect from models of social learning (Banerjee, 1992; Bikhchandani et al., 1992) and it is consistent with empirical findings such as Bursztyn et al. (2014).

individuals seem to downgrade their beliefs about efficacy when their beliefs about the share who sign up fall.

While we are able to reject the 'diminishing returns' explanation, we are unable to identify precisely what does explain the effect we observe; and it is plain that a number of explanations are possible. One possibility is that decreasing the perceived share who sign-up raises the perceived 'self-image' benefits from signing-up (Bénabou and Tirole, 2006; Falk, 2020). After all, if a person believes that hardly anybody else signs up, then signing up allows the person to tell themselves that they are a better person than nearly everybody else. A second possibility is that individuals believe that the organization is more desperate for their support when the organization sends them a signal that nobody is signing up, thereby increasing the 'social pressure' associated with the signal. Although we are unable to test the importance of these and other explanations using our data, we believe that understanding precisely what mechanisms drive this result would be an interesting task for future work.

7 Are the treatment effects purely driven by beliefs?

In principle, descriptive social norm interventions can influence individual behavior through a number of mechanisms. Most obviously, they can influence behavior by altering individual's beliefs (or point estimates) about the share that engage in that behavior. However, they can also work via other channels: for example, informing individuals about the share who do an activity can alter not just individual point estimates but also the extent to which such estimates are *salient* or at the forefront of individuals' mind (Haaland et al., 2020). In this section, we examine the extent to which the treatments change behavior (relative to the control group) purely by updating individual point estimates about prevalence.

We begin by estimating the model

$$y_i = \beta_0 + \sum_{k=1}^{6} \beta_k T_{ki} + \beta_7 b_i + \beta_8 b_i^2 + u_i \tag{6}$$

That is, we estimate the effects of the treatments on sign-ups controlling for beliefs (and also beliefs squared in light of the previous evidence that the effect of beliefs on sign-ups is non-linear). The hypothesis that the effects are *purely* driven by beliefs corresponds to the hypothesis that $\beta_k = 0$ for $k = 1, 2, ..., 6$: after controlling for beliefs, the treatments should have no effects on sign-ups. Table A9 displays the results. As can be seen, there is large (and statistically significant) effect of being allocated to the 0% group even after controlling for beliefs. As a result, an F-test strongly rejects the hypothesis that every $\beta_k = 0$ for

$k = 1, 2, ..., 6$. We take this as suggestive evidence that the effects that we observe are not purely driven by beliefs.

Having said this, it does seem that beliefs play an important role in explaining our effects. To see this, we assume (for the sake of argument) that the observed differences between the control and treatment groups are purely driven by differences in beliefs; and that beliefs influence sign-ups as specified by Equation (3). Under these assumptions, we then compute what we would expect sign-ups to be using

$$P(signup) = -0.00564b + 0.0000738b^2 \tag{7}$$

i.e., using the belief coefficients estimated using instrumental variables in Section 5. We then calculate a predicted sign-up probability for every individual using Equation (7). For every treatment group, we average over these predicted probabilities (for the individuals in the group) to obtain the expected sign-up rate for this group. As can be seen, the predicted treatment effects are quite large (see Table A8). In the 0% group, beliefs alone generate predicted effects of 2 percentage points; and in the 100% group, beliefs alone generate predicted effects of 6 percentage points (the actual effects are 5 and 8 percentage points respectively, and the predictions are just outside the standard errors of the original estimates). This indicates that a simple model in which effects are purely driven by beliefs is able to generate a sizeable proportion of the observed effects.

8 Conclusions

In this paper, we report on a descriptive social norm experiment that randomly varies the 'signal' about behavior prevalence that participants are shown. We have three main findings. First, providing information about the share who do an activity can have a sizeable effect on individual's propensity to engage in that activity. Second, the relationship between signal and behavior is convex—and plausibly even non-monotonic. Third, this convexity arises not from a convex relationship between signals and beliefs but rather from a convex relationship between beliefs and behavior. Overall, this suggests a rather complicated picture of the way in which descriptive social norms influence individual behavior.

While our findings shed light on how organizations like LLS can mobilize support, we do not claim that they will necessarily extend to all possible contexts. In some sense, our setting is rather special: for example, LLS members may care about each other and form a natural 'reference group' (Lindgren et al., 2021), which may then explain why social norms exert such

a strong influence in this environment. As a result, it is important that similar studies are conducted to provide us with a better understanding of how general these findings are—and also what precise mechanisms underpin the observed effects.

Having said this, we suspect that some of our findings may generalize to other contexts. For example, there appears to be an important qualitative difference between everybody doing an activity and most people doing it, and we conjecture that further studies will also pick this up in their analyses. Similarly, the finding that it can be useful to revise beliefs downwards may also generalize. Indeed, Wikipedia appear to have already discovered in internal tests that donations increase when they state that 'fewer than 1% of readers give' (Wikipedia, 2019), suggesting that this may be a more general phenomenon than is documented in the empirical literature (also see Linek and Traxler (2021)).

These findings are not only of academic interest but may also have implications for how policymakers and other individuals should use descriptive social norms. First, they suggest large benefits from shifting beliefs about prevalence from high to very high levels. This might be accomplished either by changing actual rates of prevalence (assuming that these are positively related to beliefs) or alternately by disclosing information about prevalence rates in settings in which prevalence is generally underestimated (see Bursztyn et al. (2020) for a recent example).

Second, our findings suggest that policymakers and organizations should be careful about revising beliefs about prevalence upwards in settings where baseline beliefs are close to zero. Indeed, they may wish to do the opposite, stressing how few do the desired behavior. This holds especially for organizations operating in low-compliance environments (e.g., an online newspaper that receives donations from a very small fraction of its readership). Such organizations can truthfully broadcast the message that hardly anybody at all is engaging in the desired activity. While our findings do not imply that such an approach will always work, they do suggest that it should be carefully considered (and tested when possible).

Finally, while our paper shows how the efficacy of norm interventions depend on the signal they convey, it is silent on the net welfare effects of different signals. Indeed, a separate line of research shows that the provision of social information can have negative effects on welfare of if the interventions, for example, induce a sufficiently large feeling of guilt (Allcott and Kessler, 2019; Butera et al., 2020). Future work should therefore explore how the types of signals used in this study influence the overall welfare of participants.

References

Akesson, J., Ashworth-Hayes, S., Hahn, R., Metcalfe, R. D., and Rasooly, I. (2020). Fatalism, beliefs, and behaviors during the covid-19 pandemic. Working Paper 27245, National Bureau of Economic Research.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics*, 95(9-10):1082–1095.

Allcott, H. and Kessler, J. B. (2019). The welfare effects of nudges: A case study of energy use social comparisons. *American Economic Journal: Applied Economics*, 11(1):236–76.

Andı, S. and Akesson, J. (2020). Nudging away false news: Evidence from a social norms experiment. *Digital Journalism*, pages 1–21.

Armantier, O., Nelson, S., Topa, G., Van der Klaauw, W., and Zafar, B. (2016). The price is right: Updating inflation expectations in a randomized price information experiment. *Review of Economics and Statistics*, 98(3):503–523.

Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3):797–817.

Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.

Bérgolo, M. L., Ceni, R., Cruces, G., Giaccobasso, M., and Perez-Truglia, R. (2017). Tax audits as scarecrows: Evidence from a large-scale field experiment. Technical report, National Bureau of Economic Research.

Bicchieri, C. and Dimant, E. (2019). Nudging with care: The risks and benefits of social information. *Public choice*, pages 1–22.

Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026.

Borsari, B. and Carey, K. B. (2003). Descriptive and injunctive norms in college drinking: a meta-analytic integration. *Journal of Studies on Alcohol*, 64(3):331–341.

Bottan, N. L. and Perez-Truglia, R. (2020). Betting on the house: Subjective expectations and market choices. Technical report, National Bureau of Economic Research.

Brent, D. A., Cook, J. H., and Olsen, S. (2015). Social comparisons, household water use, and participation in utility conservation programs: Evidence from three randomized trials. *Journal of the Association of Environmental and Resource Economists*, 2(4):597–627.

Bursztyn, L., Ederer, F., Ferman, B., and Yuchtman, N. (2014). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, 82(4):1273–1301.

Bursztyn, L., González, A. L., and Yanagizawa-Drott, D. (2020). Misperceived social norms: Women working outside the home in saudi arabia. *American Economic Review*, 110(10):2997–3029.

Butera, L., Metcalfe, R., Morrison, W., and Taubinsky, D. (2020). Measuring the welfare effects of shame and pride. Technical report, National Bureau of Economic Research.

Chen, Y., Harper, F. M., Konstan, J., and Li, S. X. (2010). Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review*, 100(4):1358–98.

Cullen, Z. and Perez-Truglia, R. (2018). How much does your boss make? the effects of salary comparisons. Technical report, National Bureau of Economic Research.

Falk, A. (2020). Facing yourself–a note on self-image. *Journal of Economic Behavior & Organization*.

Ferraro, P. J. and Price, M. K. (2013). Using nonpecuniary strategies to influence behavior: evidence from a large-scale field experiment. *Review of Economics and Statistics*, 95(1):64–73.

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

Gerber, A. S. and Rogers, T. (2009). Descriptive social norms and motivation to vote: Everybody's voting and so should you. *The Journal of Politics*, 71(1):178–191.

Haaland, I., Roth, C., and Wohlfart, J. (2020). Designing information provision experiments.

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*, 148:14–31.

Heldt, T. (2005). Conditional cooperation in the field: cross-country skiers' behavior in sweden. In *Economic Science Association, European Meeting, Alessandria, Italien, September 15-18, 2005*.

Lerner, A. P. (1944). *Economics of control: Principles of welfare economics*. Macmillan and Company Limited, New York.

Lindgren, K. P., DiBello, A. M., Peterson, K. P., and Neighbors, C. (2021). Theory-driven interventions: How social cognition can help. *The handbook of alcohol use*, pages 485–510.

Linek, M. and Traxler, C. (2021). Framing and social information nudges at wikipedia.

Perez-Truglia, R. and Cruces, G. (2017). Partisan interactions: Evidence from a field experiment in the united states. *Journal of Political Economy*, 125(4):1208–1243.

Perez-Truglia, R. and Troiano, U. (2018). Shaming tax delinquents. *Journal of Public Economics*, 167:120–137.

Roth, C. and Wohlfart, J. (2020). How do expectations about the macroeconomy affect personal expectations and behavior? *Review of Economics and Statistics*, 102(4):731–748.

Stock, J. and Yogo, M. (2005). *Asymptotic distributions of instrumental variables statistics with many instruments*, volume 6. Chapter.

Wikipedia (2019). Fundraising report: 2018-19. https://meta.wikimedia.org/w/index.php?title=Fundraising/2018-19_Report&oldid=20637485.

Appendices

A Tables and figures

Figure A1: Beliefs about sign-ups in the control group

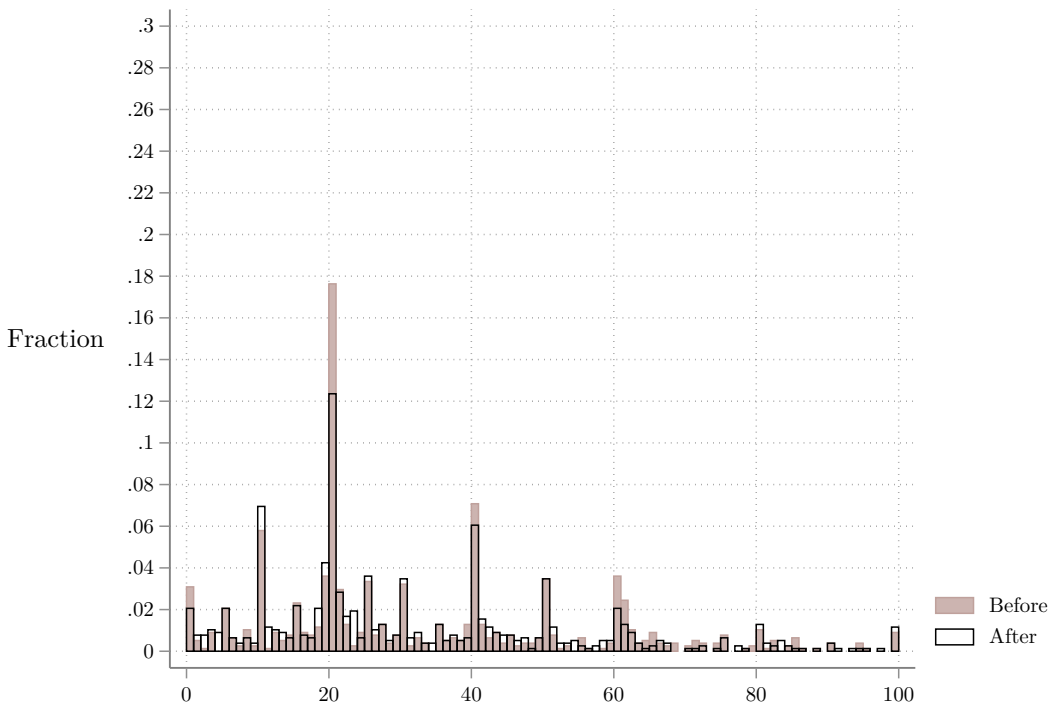


Figure A2: Beliefs about sign-ups in the 0% signal group

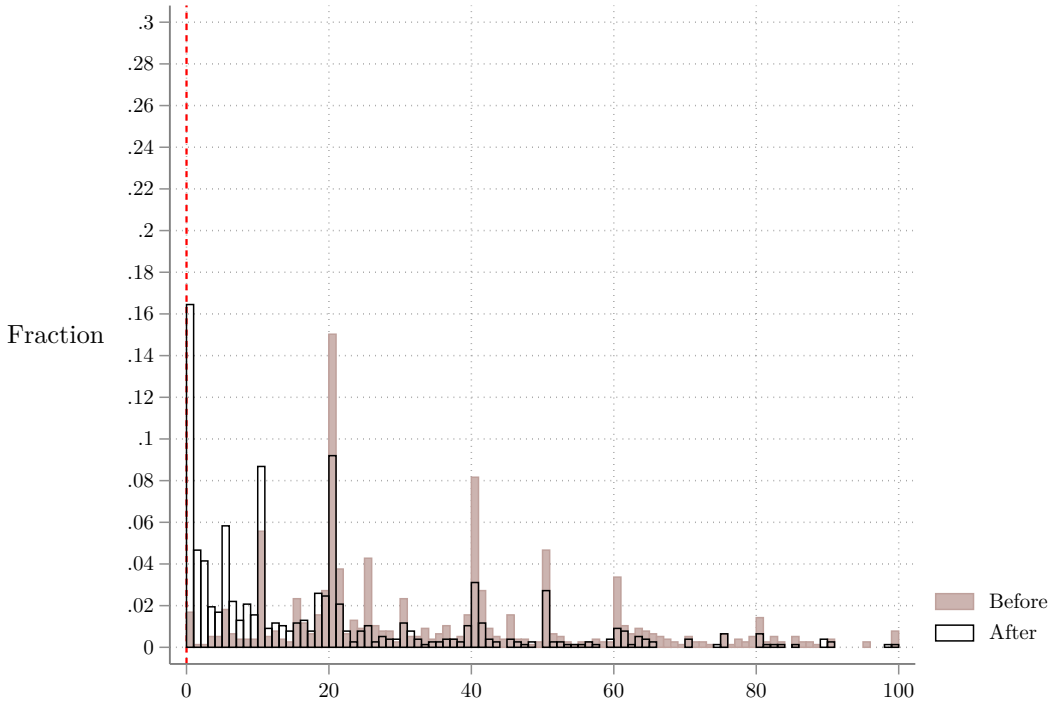


Figure A3: Beliefs about sign-ups in the 20% signal group

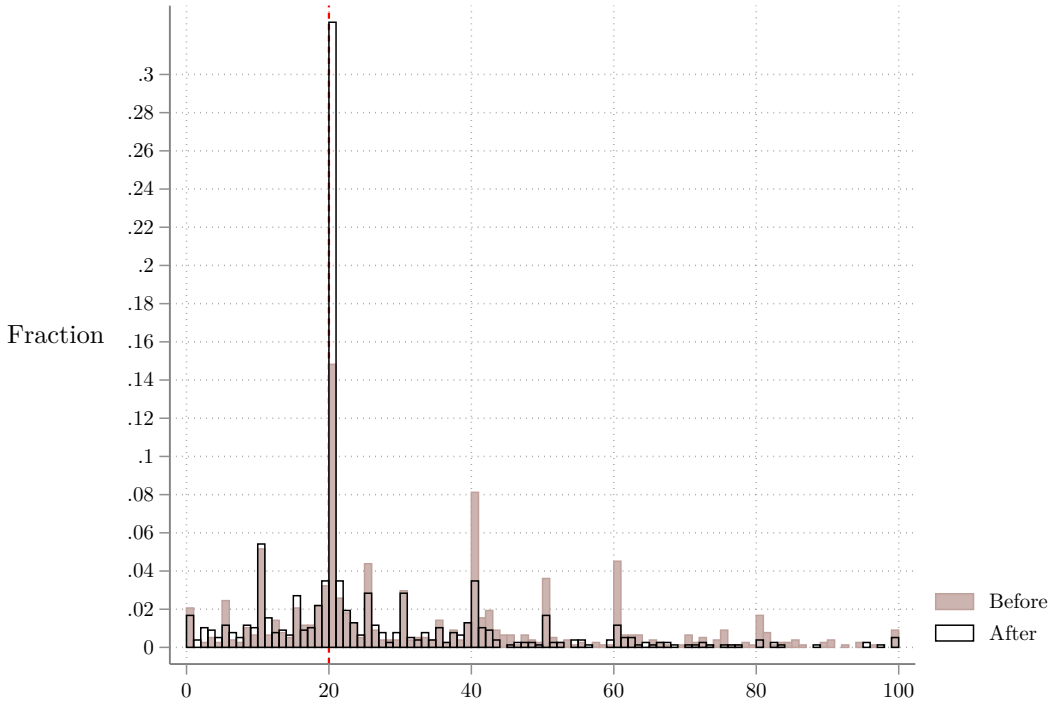


Figure A4: Beliefs about sign-ups in the 40% signal group

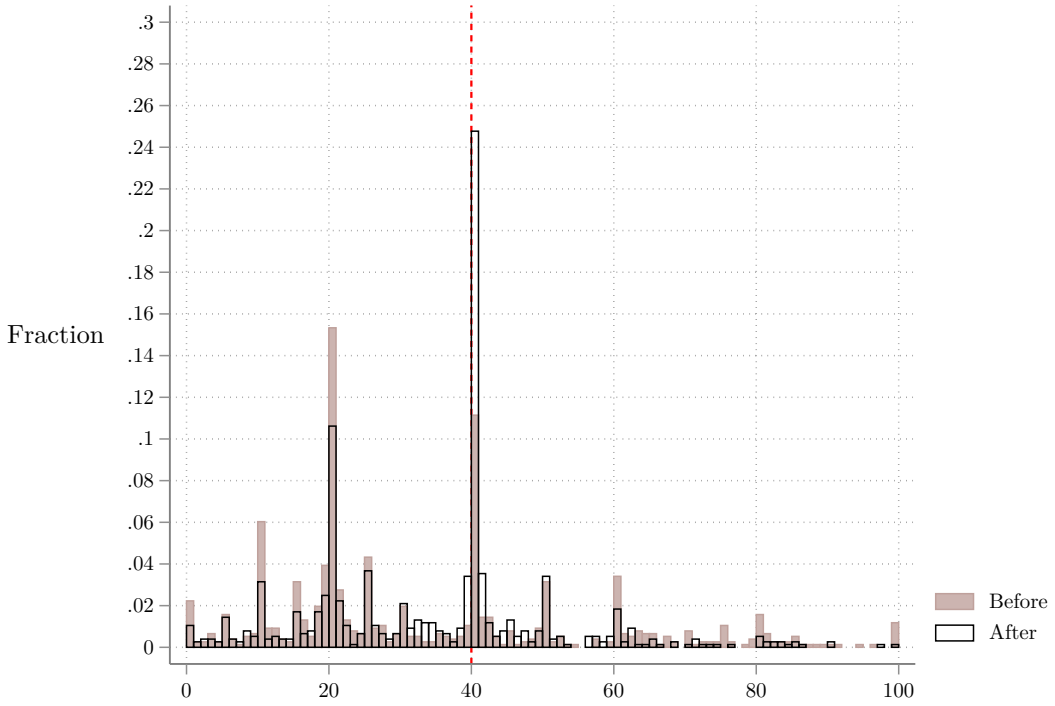


Figure A5: Beliefs about sign-ups in the 60% signal group

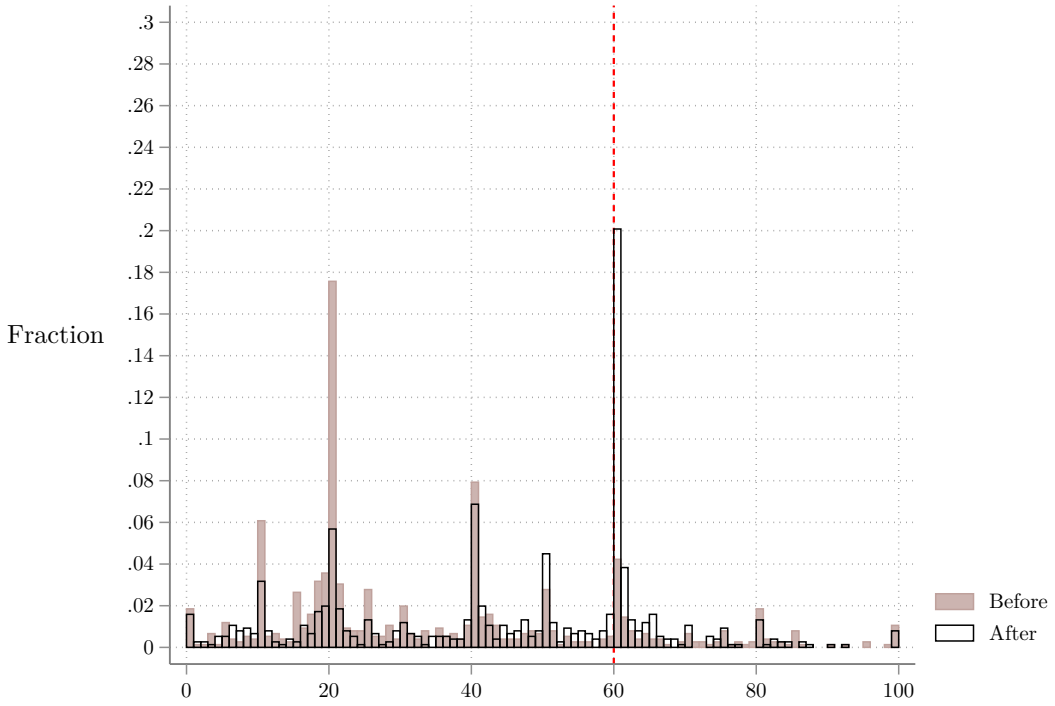


Figure A6: Beliefs about sign-ups in the 80% signal group

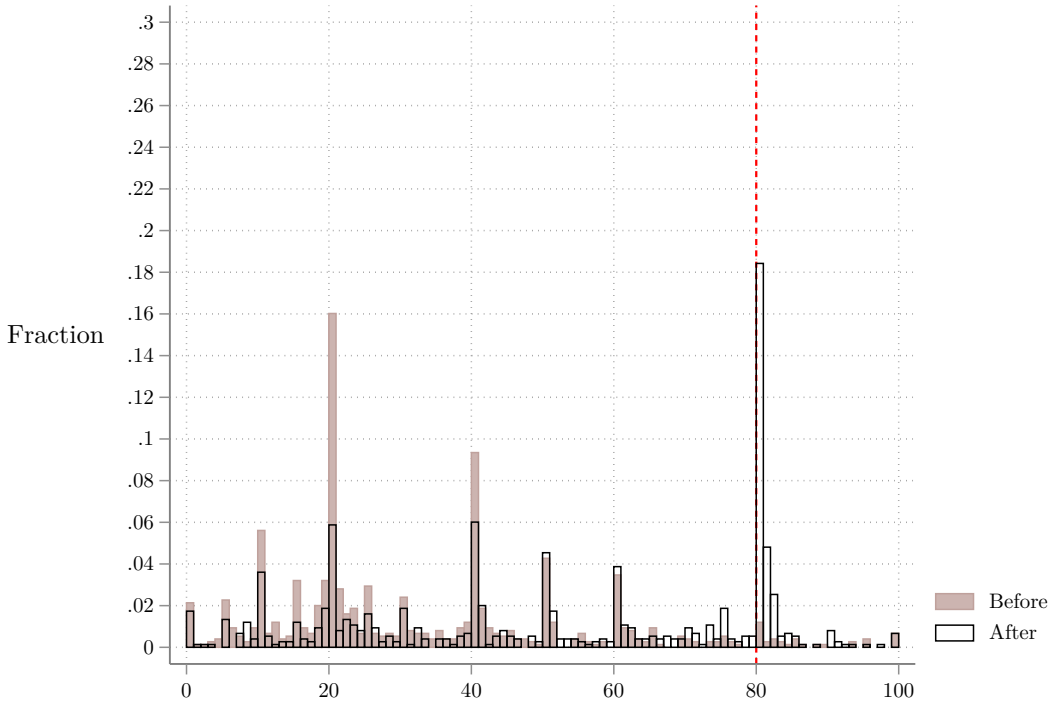


Figure A7: Beliefs about sign-ups in the 100% signal group

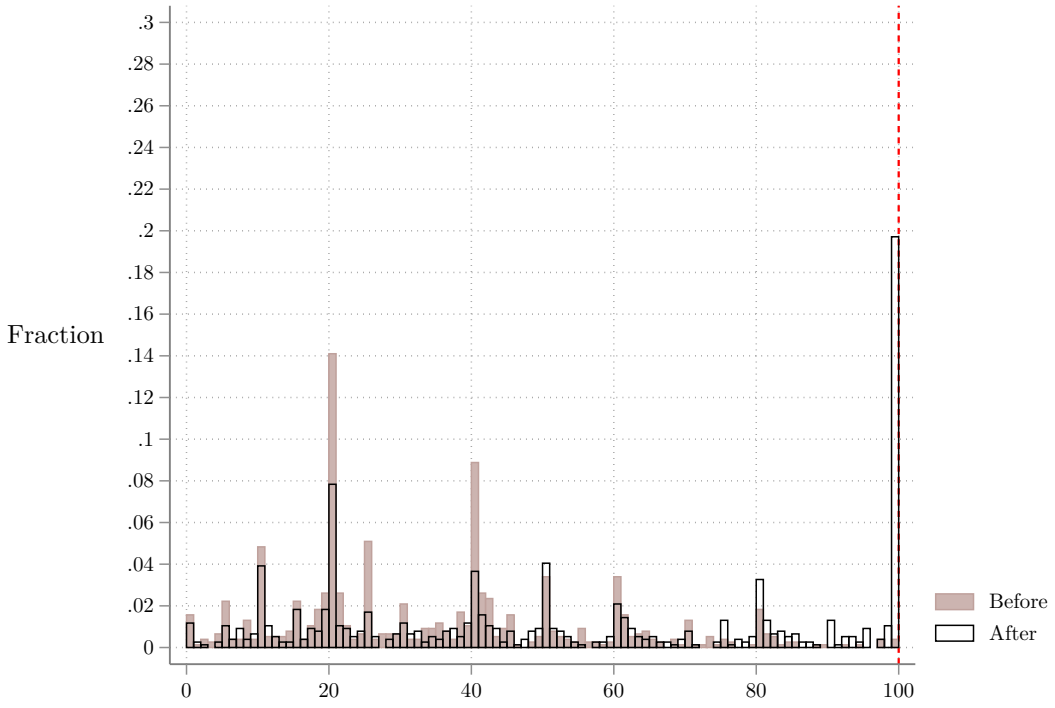


Table A1: Balance table

	Control	Signal 0%	Signal 20%	Signal 40%	Signal 60%	Signal 80%	Signal 100%	*p*-value
Age (years)	64.0	63.8	64.1	64.2	64.4	64.4	63.6	0.820
Familiar with LLS advocacy (binary)	0.663	0.703	0.688	0.702	0.724	0.673	0.689	0.177
BSc or more (binary)	0.495	0.509	0.512	0.491	0.484	0.529	0.504	0.740
Pre-treatment beliefs	32.7	34.3	33.2	33.4	33.8	32.1	34.2	0.419
Pre-treatment confidence in beliefs	1.995	1.970	1.912	1.931	1.917	1.917	1.957	0.250
n	777	772	776	763	757	749	766	

Notes. This tables presents averages for all pre-treatment variables. We are missing some values for age and education, and so we present the averages for those who responded to those questions. The right-most column (labeled '*p*-value') presents *p*-values for joint orthogonality tests of treatment arms.

Table A2: Treatment effect on share that sign up

	(1) Effect on sign-up shares	(2) Effect on sign-up shares
0% Signal	0.0442**	0.0348
	(0.0218)	(0.0227)
20% Signal	0.000285	-0.00177
	(0.0211)	(0.0221)
40% Signal	0.00799	-0.00200
	(0.0213)	(0.0221)
60% Signal	0.0243	0.00981
	(0.0216)	(0.0225)
80% Signal	0.0336	0.0363
	(0.0218)	(0.0232)
100% Signal	0.0828***	0.0761***
	(0.0223)	(0.0234)
Constant	0.221***	0.166***
	(0.0149)	(0.0445)
Controls	No	Yes
Observations	5,360	4,368
R^2	0.004	0.068

Notes. In columns (1) and (2), we regress whether participants signed up (binary) on treatment assignment. The control group is the omitted category. In column (2), we also control for: age, education, pre-treatment beliefs, pre-treatment confidence in beliefs, and familiarity with LLS advocacy efforts. The sample is smaller for the second regression because we do not have demographic information for all individuals.

Table A3: Treatment effect on confidence in beliefs

	(1) Confidence in beliefs (Likert)	(2) Confident or very confident (binary)	(3) Confidence in beliefs (Likert)	(4) Confident or very confident (binary)
0% signal			0.185***	0.109***
			(0.0421)	(0.0210)
20% signal	-0.0330	-0.0130	0.152***	0.0956***
	(0.0464)	(0.0226)	(0.0421)	(0.0208)
40% signal	-0.00213	-0.0138	0.183***	0.0948***
	(0.0453)	(0.0227)	(0.0409)	(0.0209)
60% signal	0.0347	0.0108	0.220***	0.119***
	(0.0450)	(0.0230)	(0.0407)	(0.0213)
80% signal	0.0140	-0.0182	0.199***	0.0904***
	(0.0452)	(0.0227)	(0.0408)	(0.0209)
100% signal	0.0364	0.0152	0.221***	0.124***
	(0.0468)	(0.0230)	(0.0426)	(0.0212)
Constant	2.153***	0.277***	1.968***	0.169***
	(0.0327)	(0.0161)	(0.0264)	(0.0134)
Observations	4,583	4,583	5,360	5,360
R^2	0.001	0.001	0.007	0.008

Notes. This table presents the results from four regressions. In columns (1) and (3), we regress the degree to which participants are confident in their beliefs (measured on a five-point Likert scale) on treatment assignment. In columns (2) and (4), we regress whether participants are confident or very confident in their beliefs (binary outcome) on treatment assignment. In columns (1) and (2) we omit the control group (Signal = 0% is the reference group). No observations are omitted in columns (3) and (4), and the control group is the reference group. Robust standard errors in parentheses (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$).

Table A4: Treatment effect on beliefs about the share that sign up

	(1) Beliefs about sign-ups	(2) Beliefs about sign-ups
0% Signal	-13.15***	-13.25***
	(1.064)	(0.832)
20% Signal	-6.363***	-5.266***
	(0.984)	(0.738)
40% Signal	2.489**	3.750***
	(0.990)	(0.778)
60% Signal	12.21***	12.04***
	(1.106)	(0.971)
80% Signal	19.45***	21.01***
	(1.262)	(1.182)
100% Signal	22.65***	22.79***
	(1.428)	(1.364)
Constant	31.37***	16.06***
	(0.782)	(2.091)
Controls	No	Yes
Observations	5,360	4,368
R^2	0.222	0.398

Notes. This table presents the results of two regressions. In columns (1) and (2), we regress whether participants' beliefs about the share that sign up on treatment assignment. The control group is the omitted category. In column (2), we also add the following controls to the regression: age, education, pre-treatment beliefs, pre-treatment confidence in beliefs, and familiarity with LLS advocacy efforts. The sample is smaller for the second regression because we do not have demographic information for all individuals. Robust standard errors in parentheses (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$).

Table A5: Effects of beliefs on sign-ups (1)

	(1)	(2)	(3)	(4)	(5)
	% sign up	% sign up	% sign up	% sign up	% sign up
Belief % sign up	0.00104**	-0.00564***	-0.00632	-0.0197*	-0.0199
	(0.000492)	(0.00204)	(0.00490)	(0.0108)	(0.0268)
" ˆ2		7.38e-05***	9.21e-05	0.000763	0.000783
		(2.19e-05)	(0.000123)	(0.000500)	(0.00194)
" ˆ2, ˆ3			-1.26e-07	-1.11e-05	-1.16e-05
			(8.36e-07)	(7.99e-06)	(5.23e-05)
" ˆ2, ˆ3, ˆ4				5.59e-08	6.22e-08
				(4.06e-08)	(5.90e-07)
" ˆ2, ˆ3, ˆ4, ˆ5					-0
					(2.35e-09)
Constant	0.215***	0.309***	0.314***	0.365***	0.365***
	(0.0193)	(0.0343)	(0.0474)	(0.0603)	(0.0829)
Observations	4,583	4,583	4,583	4,583	4,583
R^2	0.021	0.013	0.014	0.015	0.015

Notes. Robust standard errors in parentheses (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$). In this table, we present the results from five instrumental variable regressions. The outcome is always the share that sign up. In each case, we omit the control group, and use the treatment assignment to instrument for beliefs. In the first model, we only instrument for beliefs. In the other four models, we also instrument for polynomial belief terms. We do not use any control variables in these regressions.

Table A6: Effects of beliefs on sign-ups (2)

	(1)	(2)	(3)
	% sign up	% sign up	% sign up
Belief % sign up	-0.00754*	-0.00582	
	(0.00387)	(0.0227)	
" if belief $\geq$ 20	0.00969		
	(0.00694)		
" if belief $\geq$ 40	-0.000818		
	(0.00744)		
" if belief $\geq$ 60	-0.000361		
	(0.00779)		
" if belief $\geq$ 80	0.0108		
	(0.00864)		
" = 0		0.0521	
		(1.032)	
" if belief $\geq$ 25		0.0101	
		(0.0302)	
" if belief $\geq$ 50		-0.00628	
		(0.0120)	
" if belief $\geq$ 75		0.0134	
		(0.00899)	
Belief > 20 & $\leq$ 40			-0.0929
			(0.0631)
Belief > 40 & $\leq$ 60			0.0181
			(0.0650)
Belief > 60 & $\leq$ 80			-0.0609
			(0.0839)
Belief > 80 & $\leq$ 100			0.213***
			(0.0723)
Constant	0.337***	0.312	0.262***
	(0.0497)	(0.427)	(0.0228)
Observations	4,583	4,583	4,583
R^2	0.014	0.012	0.001

Notes. Robust standard errors in parentheses (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$). In this table, we present the results from three instrumental variable regressions. The outcome is always the share that sign up. In each case, we omit the control group, and use the treatment assignment to instrument for beliefs.

Table A7: Effects of beliefs on perceived efficacy

	(1) Perceived efficacy	(2) Perceived efficacy	(3) Perceived efficacy	(4) Perceived efficacy	(5) Perceived efficacy
Belief % sign up	0.00132**	0.00236	0.0122**	0.00234	0.0150
	(0.000513)	(0.00214)	(0.00517)	(0.0116)	(0.0292)
" ˆ2		-1.15e-05	-0.000277**	0.000215	-0.000754
		(2.31e-05)	(0.000130)	(0.000540)	(0.00212)
" ˆ2, ˆ3			1.83e-06**	-6.22e-06	2.05e-05
			(8.88e-07)	(8.62e-06)	(5.73e-05)
" ˆ2, ˆ3, ˆ4				4.11e-08	-2.62e-07
				(4.38e-08)	(6.45e-07)
" ˆ2, ˆ3, ˆ4, ˆ5					1.21e-09
					(2.57e-09)
Constant	0.261***	0.246***	0.175***	0.213***	0.183**
	(0.0202)	(0.0352)	(0.0487)	(0.0624)	(0.0878)
Observations	4,583	4,583	4,583	4,583	4,583
R^2	0.015	0.013	0.011	0.010	0.010

Notes. Robust standard errors in parentheses (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$). In this table, we present the results from five instrumental variable regressions. The outcome is always the share that perceived efficacy of LLS advocacy. In each case, we omit the control group, and use the treatment assignment to instrument for beliefs. In the first model, we only instrument for beliefs. In the other four models, we also instrument for polynomial belief terms. We do not use any control variables in these regressions.

Table A8: Predicted vs. actual sign up rates

	(1) Actual % sign up	(2) Predicted % sign up
Signal = 0%	0.0442**	0.0206***
	(0.0218)	(0.00224)
Signal = 20%	0.000285	-0.00518**
	(0.0211)	(0.00208)
Signal = 40%	0.00799	-0.0164***
	(0.0213)	(0.00206)
Signal = 60%	0.0243	-0.00227
	(0.0216)	(0.00219)
Signal = 80%	0.0336	0.0273***
	(0.0218)	(0.00270)
Signal = 100%	0.0828***	0.0604***
	(0.0223)	(0.00420)
Constant	0.221***	0.240***
	(0.0149)	(0.00167)
Observations	5,360	5,360
R^2	0.004	0.155

Notes. Robust standard errors in parentheses (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$). In column (1), we regress whether participants signed up (binary) on treatment assignment. The control group is the omitted category. In column (2), our outcome is the predicted value for whether participants signed up or not (using the model specified in column (2) of Table A5).

Table A9: Treatment effects after controlling for beliefs

	(1)	(2)
	% sign up	% sign up
Signal = 0%	0.0991***	0.0985***
	(0.0211)	(0.0214)
Signal = 20%	0.0269	0.0268
	(0.0204)	(0.0204)
Signal = 40%	-0.00240	-0.00210
	(0.0208)	(0.0210)
Signal = 60%	-0.0266	-0.0264
	(0.0211)	(0.0212)
Signal = 80%	-0.0476**	-0.0476**
	(0.0219)	(0.0219)
Signal = 100%	-0.0117	-0.0123
	(0.0219)	(0.0221)
Post-treatment beliefs	0.00418***	0.00407***
	(0.000256)	(0.000809)
Post-treatment beliefs (squared)		1.16e-06
		(8.80e-06)
Constant	0.0904***	0.0920***
	(0.0154)	(0.0185)
Observations	5,360	5,360
R^2	0.053	0.053

Notes. Robust standard errors in parentheses (*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$). In columns (1) and (2), we regress whether participants signed up (binary) on treatment assignment. The control group is the omitted category. In both cases, we control for post-treatment beliefs. In column (2), we also control for a squared belief term.