European Economic Review 134 (2021) 103711

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/euroecorev

When a nudge is (not) enough: Experiments on social information and incentives☆

Jingnan (Cecilia) Chen[a], Miguel A. Fonseca[a,b,*], Shaun B. Grimshaw[a]

[a] Economics Department, Business School, University of Exeter, Exeter EX4 4PU, UK
[b] NIPE, Universidade do Minho, Portugal

ARTICLE INFO

Article history:
Received 23 June 2020
Revised 26 January 2021
Accepted 5 March 2021
Available online 16 March 2021

JEL classification:
C93
D01
D91

Keywords:
Field experiment
Financial incentives
Social information
Cooperation
Public goods
Behavior change

ABSTRACT

Financial incentives and information nudges are two of the most widely used behaviour change interventions. However, we do not yet fully understand how incentives and social information interact. We report two experiments examining how incentives and social information interact to induce behavior change. In the first experiment, the behavior of interest is punctuality in the field; in the second, we examine cooperation in a large-N prisoners' dilemma in the lab. In both experiments participants valued good behavior and believed others also valued it, yet only a minority behaved well. We find that incentives work in both environments, while information nudges were only effective in the prisoners' dilemma. Incentives complement information nudges only in the prisoners' dilemma. Our experimental design also allows us to distinguish between intrinsically motivated and unmotivated subjects: the former respond to treatment manipulations very differently to the latter, both behaviourally and in their beliefs about others' behavior.

1. Introduction

People are often reluctant to engage in socially desirable behaviors despite recognizing their importance. As a result, encouraging people to do so is part and parcel of public policy: tax agencies urge taxpayers to pay their taxes on time; health authorities want their citizens to adopt healthier lifestyles.[1] To achieve the desired behavioral change, economists typically prescribe financial incentives. More recently, economists have been examining measures that target non-economic motives. One such example is providing information on others' decisions: descriptive social information as a social nudge. The popularity of nudges comes from their low cost and their ability to retain freedom of choice (Thaler and Sunstein, 2008).

☆ Financial support from ESRC grant ES/N00762X/1 and the University of Exeter Business School is gratefully acknowledged. We thank two anonymous reviewers for invaluable comments, Anthony Heyes, Daniel Houser, Julian Jamison, Daniele Nosenzo, David Reinstein and Roberto Weber, as well as participants at seminars at George Mason, Exeter, IQTE, Renmin and Tsinghua universities, as well as the 2017 North American ESA conference for their valuable comments and suggestions. The usual disclaimer applies.

* Corresponding author at: Economics Department, Business School, University of Exeter, Exeter EX4 4PU, UK.
E-mail addresses: j.chen2@exeter.ac.uk (J. Chen), m.a.fonseca@exeter.ac.uk (M.A. Fonseca), s.b.grimshaw@exeter.ac.uk (S.B. Grimshaw).

[1] The late payment of taxes is an important component of the tax gap: in 2001, it accounted for 16% of all uncollected tax in the US (Slemrod, 2007). The rise in obesity rates in the US has led to calls for research on how to promote healthy lifestyle choices (Health, 2007).

https://doi.org/10.1016/j.euroecorev.2021.103711
0014-2921/

In this paper we investigate the effectiveness of two types of interventions – financial incentives and social information – under two distinct decision contexts. We not only study the main behavioral effect of incentives and social information, but also how incentives may interact with social information to shape behavior. We are particularly interested in understanding the circumstances under which incentives and social norm interventions may complement or undermine each other and more importantly shed light on the underlying mechanism that drives the changes in behavior.

To this effect, we conducted two studies with the same experimental design: one in the field, the other in the lab. Our field experiment examines punctuality, in particular punctual attendance by undergraduate students to two experimental sessions on consecutive days that started at 9am. Importantly, participants were not aware that their arrival time was the subject of interest, therefore making it a natural field experiment. Punctuality is a universally acknowledged and valued mode of behavior and makes it the ideal subject for our main research question: what is the best lever to change individual behavior when there is a common understanding of what is the right thing to do?

The lab study was done online with a non-student sample. The behavior of interest was cooperation in a large-N, one-shot prisoners' dilemma, where participants chose whether or not to cooperate. The second study complements the first by examining an environment in which the externality imposed on others by bad behavior is more salient and financially quantifiable. Many public policy interventions target behaviors that have negative economic externalities at the population level. For example, tax non-compliance, drink driving, and unhealthy diets. All these above-mentioned examples create burdens to public services, which result in either worse public services or higher taxes. A large N social dilemma captures the essence of these issues well. Despite their obvious differences, the behaviors we investigate in the field and lab study share a crucial common feature: they are both (punctuality and cooperation) highly valued good behavior, yet few chose to engage in it.

In both studies, behavior on the first day was used to calibrate individual baseline behavior. On day two, subjects were randomly assigned to one of 9 treatments in a 3×3 between-subjects factorial design, which varied along two dimensions. The first dimension was an information manipulation: depending on the condition, prior to making their decisions in the second day (i.e. to arrive on time or to cooperate), subjects were *truthfully* informed that most subjects with whom they would interact had behaved appropriately, or inappropriately.[2] In the control condition they received no information. The second dimension was a financial incentive manipulation: depending on the condition, prior to the start of the second day session, subjects were informed that they would receive a reward if they behaved appropriately, or a penalty if they behaved inappropriately; in the control condition, they did not receive any financial incentive.

We find that financial incentives were effective in changing behavior in both decision contexts and across behavioral types. In contrast, social information nudge and its interaction with incentives were mainly effective in the social dilemma context. Intrinsically motivated types differed from unmotivated ones in their responses to treatment interventions, especially to information nudge interventions. In both studies, changes in behavior are best explained by a combination of behavioral types (motivated vs. unmotivated types as proxied by persistent individual behavior in the absence of any intervention), injunctive norms pertaining to the importance of engaging in the socially appropriate behavior, and first-order beliefs about the behavior of others. In this sense, our findings complement that of Kölle et al. (2020) who find that incentives can strengthen or undermine social norms in a voter registration context.

Further, we find that information nudges were most effective in the cooperation experiment, where the externality inherent to cooperation was most transparent. There, social information likely served as an aid to conditional cooperators, who would otherwise choose not to cooperate. In the punctuality environment, however, where the externality from good behavior was more diffuse, the income effect was necessary to sway the behavior of those for whom it was too costly to make it on time. Finally, we find strong evidence of heterogeneity in behavioral and belief responses to treatment. This is a critical dimension of behavioral change, one which should be further explored in future research.

Social norm based interventions have become increasingly popular over the last three decades starting with Cialdini et al. (1990), and have been used in a variety of contexts: female labor force participation (Bursztyn et al., 2020), tax compliance (Blumenthal et al., 2001; Hallsworth et al., 2017; Bott et al., 2020), medication prescriptions (Hallsworth et al., 2016), and household energy consumption (Allcott, 2011; Allcott and Kessler, 2019). Despite the increased use of social information nudges in public policy, the empirical evidence on the effectiveness of social information in changing behaviours is mixed (Bicchieri and Dimant, 2019; Richter et al., 2018), ranging from economically large effects (Bursztyn et al., 2020; Hallsworth et al., 2017) to precisely estimated zero effects (Silva and John, 2017; Dimant et al., 2020). Understanding what determines the effect of a behavioral intervention, and how sensitive intervention outcomes are to changes in the choice context is essential for policy design.

Meanwhile, there is a growing recognition among economists that the use of extrinsic motivators can be counterproductive (Gneezy and Rustichini, 2000a), and internal reward systems are likely to be at the heart of such behavioral responses (Bernheim, 1994; Bénabou and Tirole, 2006). Social norms have been theorized as intrinsic motivators of behavior (Cialdini and Trost, 1998) and are perceived by individuals as such (Nolan et al., 2008). It is therefore theoretically and practically relevant to understand the behavioral interplay between social norms and incentives.

[2] This introduced a constraint in the random assignment to treatment such that the distribution of day one behavior was not the same in expectation or realization across treatments.

Despite of its burgeoning importance, little has been done on the impact of combining social information interventions with financial incentives on behavior. The literature is very recent and the evidence is rather mixed. Burtch et al. (2018) examine this question in the context of online marketplace reviews and find that combining social norms with incentives lead to improvements in the length and quality of reviews relative to either condition in isolation. Thorndike et al. (2016) find a similar result in the context of healthy eating choices in a hospital cafeteria. In contrast, Kullgren et al. (2014) find that combining norms and incentives did not result in increase in outdoor exercise among older adults. Pellerano et al. (2017) find that adding economic incentives to normative interventions actually undermines the positive effect of norms on energy consumption. Knowing when to use incentives or social information intervention and when to combine those two measures are crucial to policy makers, given the ubiquity of both types of interventions in contemporary societies. We contribute to this literature by reporting on two experimental studies that systematically examined how social information, financial incentives and their interactions to promote behavioral change.

An important feature of our experimental design is that we can distinguish between participants who are intrinsically motivated to behave well, and those who are not. We make this distinction through behavior on day one of the experiment, where (especially in our field study) participants were unaware that we were examining their behavior in any specific way. We can therefore directly identify and exploit individual heterogeneity in intrinsic motivation to behave, something most studies in the area of behavior change have not done to date. Hallsworth et al. (2017), Allcott and Kessler (2019) and Hauser et al. (2019) are notable exceptions. Hallsworth et al. (2017) condition their analysis of the effect of different norm conditions on tax debt repayment on the degree of debt. They find some evidence of heterogeneity in response to treatment. Hauser et al. (2019) go a step further and use machine learning methods to predict risk of fraud in unemployment state benefit claims in a US state. They find response to norm interventions were positive and significant for claimants with a high fraud risk assessment, but not for those with low fraud risk; for some subsets of applicants, some nudges had a negative effect. Allcott and Kessler (2019) elicit WTP for Home Energy Reports (a social information nudge) from home owners who have received HERs and estimate the welfare benefits of the nudge. They find significant heterogeneity in the sample: some homeowners have negative WTP (they would prefer not to be nudged) while others have very high WTP. The overall welfare effect is positive. However, the authors argue for a machine learning approach to identify those consumers for which the welfare gain of the nudge is positive. Such an approach doubles the welfare gains from the policy. We are the first study that directly examines individual heterogeneity and are able to directly estimate the behavioral response to financial incentives and social norm nudges.

The remainder of the paper is organized as follows: Section 2 outlines a simple theoretical framework and the hypotheses that underpin the two studies. Section 3 presents the first experiment and its results. Section 4 outlines the second experiment and its the results. Section 5 provides a discussion of the findings and concludes the paper.

2. Theory and hypotheses

In this section we outline a simple theoretical framework to analyze how financial incentives and social information may affect behavior and to generate hypotheses. We consider a setting with one principal and n agents, all of whom are risk neutral. The principal wishes agents to engage in a particular course of action. This could be a tax agency wanting taxpayers to file their taxes by a particular date; a school authority that wants its teachers to be present in the classroom; a public health authority that wants patients to take their medication regularly. We will denote this as the 'appropriate action', or 'good behavior' – we will use both expressions interchangeably.

We assume that there is no monitoring problem: the principal can detect deviations at zero cost. This is consistent with the empirical literature on incentives in which the incentive structure has a built-in monitoring mechanism (see Gneezy et al., 2011). We also assume agents do not incur reputational or self-image concerns (Bénabou and Tirole, 2006). Subjects cannot observe each others' actions; they could have image or reputation concerns with respect to the beliefs the experimenters hold about their type, but given that subjects were not physically co-present with the experimenters—both experiments were done anonymously online—these behavioral mechanisms should not play a major role.

2.1. The environment

We consider a one-period, one-shot environment. Agents have a choice between undertaking the appropriate action or not. Denote I as an indicator function for an agent undertaking the appropriate action. If agent i undertakes the appropriate action, she incurs a cost $c_i \geq 0$; if that agent does not do the appropriate action, she experiences a private benefit $B_i \geq 0$. Following Krupka and Weber (2013), we assume that if subject i chooses the appropriate action, she derives some utility from abiding by the socially appropriate custom, $\phi_i S$; the parameter $\phi_i \in [0, 1]$ denotes the extent to which subject cares about behaving in the socially prescribed manner.[3] The principal can employ financial incentives to induce unmotivated types to take the appropriate action. The principal may wish to either punish bad behavior by levying a fine F, or giving a reward R for good behavior.

[3] We could have modelled the injunctive norm as a penalty from arriving late, since deviating from an injunctive norm ought to lead to a disutility from social censure, but the results would be unaltered. We chose this particular approach to be consistent with Krupka and Weber (2013).

While agents are influenced by internal costs and benefits, such motivations could be undermined in a population which generally behaves differently to how they do. Agents may care about what the prevailing behavior is among the population with whom they interact. Consequently, the more people in their group fail to take the appropriate action, the greater their desire to do the same. We operationalize the effect of social information in a similar manner as the literature on social customs Akerlof (1980) and Myles and Naylor (1996) by assuming each subject derives utility from taking the appropriate action, $\gamma_i N(\mu)$, which is a non-decreasing function of μ, the belief about the proportion of other subjects who are taking the appropriate action. The parameter $\gamma_i \in [0, 1]$ is the degree to which subject i cares about following the crowd. Note that we model social information only as a coordination device, rather than a signalling medium.

Following these considerations, we propose the following utility function for subject i:

$$U_i = (1 - l)(\phi_i S - c_i) + l(B_i) + (1 - l)R - lF + \gamma_i N(\mu) \tag{1}$$

If we assume incentives and social information are absent (i.e. $(1 - l)R = F = \gamma_i N(\mu) = 0$), we can derive two types of agents: those have strong innate preference for acting appropriately, for whom $\phi_i S - c_i \geq B_i$, which we will call "motivated" types. There are also those who require additional incentives to behave appropriately, for whom $\phi_i S - c_i < B_i$, who we will denote as "unmotivated" types.

In the absence of social information, the principal can employ financial incentives to induce unmotivated types to take the appropriate action. In this case, as long as $R + \phi_i S - c_i > B_i$, or if $\phi_i S - c_i > B_i - F$, then agent i will have an incentive to take the appropriate action. In other words, in the absence of social information, some proportion of unmotivated types will take the appropriate action if exposed to a sufficiently high reward or a fine (Gneezy and Rustichini, 2000b). In contrast, motivated types will be unaffected by either a reward or a fine. This leads to the first hypothesis.

Hypothesis 1. Unmotivated types will increase their relative frequency of good behavior, when sufficiently high financial incentives are introduced.

We now consider the case where financial incentives are not available to the principal but the principal can use social information. An agent will choose to take the appropriate action if $\phi_i S + \gamma_i N(\mu) - c_i \geq B_i$ or choose not to do so if $\phi_i S + \gamma_i N(\mu) - c_i < B_i$. It is possible to evaluate the critical proportion of agents taking the appropriate action required for an unmotivated subject to do the same. If there exists a value of μ^* such that

$$\phi_i S + \gamma_i N(\mu^*) - c_i = B_i \tag{2}$$

then, since the left-hand side of (2) is monotonically increasing in μ, an agent will take the appropriate action if $\mu \geq \mu^*$, but will not do so if $\mu < \mu^*$.[4]

Now consider an agent for whom there is a μ^* such that Eq. (2) is satisfied. From that equality we can calculate the effect of changes in the key parameters on μ^*. In particular,

$$\frac{\mathrm{d}\mu^*}{\mathrm{d}c_i} = \frac{\mathrm{d}\mu^*}{\mathrm{d}B_i} = \frac{1}{\gamma_i N'} > 0 \tag{3}$$

Eq. (3) shows that the higher the cost of effort to take the appropriate action, or the higher the private benefit B_i from not taking the appropriate action, the higher the threshold μ^* must be to induce behavior change. In other words, in the absence of financial incentives, some unmotivated agents may take the appropriate action if they believe that enough of their fellow agents are doing so. Conversely, some motivated agents may choose not to take the appropriate action if they believe that most people in the population are doing the same. This leads to the next hypotheses.

Hypothesis 2. Unmotivated types will reduce their relative frequency of bad behavior, when social information about prevalent good behavior is introduced.

Hypothesis 3. Motivated types will increase their frequency of bad behavior, when social information about prevalent bad behavior is introduced.

We conclude by extending the analysis of the effect of social information to incorporate either fines or rewards. Eq. (2) becomes

$$\phi_i S + \gamma_i N(\mu^*) - c_i + (1 - l)R = B_i - lF \tag{4}$$

The effect of changes in F and R on μ^* is given by the following:

$$\frac{\mathrm{d}\mu^*}{\mathrm{d}F} = \frac{\mathrm{d}\mu^*}{\mathrm{d}R} = -\frac{1}{\gamma_i N'} < 0 \tag{5}$$

Fines and rewards interact with social information as one would expect: the higher the fine or the reward, the lower μ^* can be. In other words, financial incentives enhance the effectiveness of social information.

[4] Note that if $\phi_i S + \gamma_i N(1) - c_i < B_i$, then agent i will never take the appropriate action irrespective of μ, while if $\phi_i S + \gamma_i N(0) - c_i \geq B_i$, then agent i will always do so.

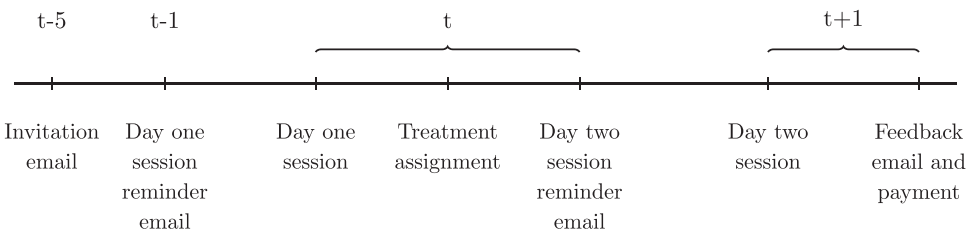

Fig. 1. Timeline of Experiment.

Hypothesis 4. Unmotivated types will decrease their frequency of bad behavior, when both incentives and social information about good behavior are introduced.

Financial incentives can interact with social information in an additional way. If the principal is better informed than each individual agent about the behavior in the population of agents, by imposing a financial incentive on a particular mode of behavior, the principal is revealing information about the true prevalence of that behavior in the population. Increasing the reward or the fine could shift μ downwards.[5] This in turn could lead to a decrease in the rate of appropriate behavior if the negative change in $N(\mu)$ is larger than the change in F or R in Eq. (4). Galbiati et al. (2013) find evidence of this mechanism in minimum-effort games; Van Der Weele (2012) provides a model of signalling through sanctions in the context of public good provision. This leads to the final hypothesis.

Hypothesis 5. Motivated types will increase their frequency of bad behavior, when both incentives and social information about the prevalent bad behavior are introduced.

3. Study I

3.1. Experimental design

The experiment was conducted between January and November 2016. We invited 800 undergraduate students using the ORSEE (Greiner, 2015) recruitment system at the FEELE lab at the University of Exeter to participate in an online economic experiment over two consecutive days — for consistency, all sessions were conducted on Tuesdays and Wednesdays. Of those invited, 739 completed both days (40% male, mean age 19.70).[6] All sessions were scheduled to start at exactly 9am. During both experimental sessions, the participants played a set of incentivized economic games.[7] The experimental manipulation was the information subjects received about the incentives for timely participation and/or punctuality norms in the subject pool. Fig. 1 illustrates the timeline of the experiment.

In the initial invitation email (see Experimental Materials Section A1), the participants were informed that they would receive a show-up fee of £3 for each day of the experimental session in addition to their experimental earnings and that the experimental payments would be transferred automatically to their bank accounts using the university's payments system upon the conclusion of the second session. The email text specified the start and end time of both sessions.

In the FEELE lab, there is an expectation on the part of experimenters that subjects turn up to sessions at the appointed start time.[8] Importantly, other than the fact that the experimental sessions were to be conducted online at 9am, a particularly early time, nothing else about the initial invitation email would lead subjects to perceive the experiment differently from any other experiment conducted in the FEELE lab. We also note that classes at the university start at 35 min past the hour mark (the first class of any day start at 8:35am). As such, there should be no class-related reason for a subject who signed up for the experiment to be late.

The evening before the day-one session, we sent a reminder email to participants. In this email, we reminded the subjects of the session time, and informed them of the web address that they should use, as well as their individual username and password to access the website in which they would make their decisions. At the end of the day-one session, partici-

[5] It could potentially leave μ unchanged if $\mu = 0$.

[6] Attrition was very low, and not did not vary systematically across treatments (chi2(8)= 5.365, p=0.718). For full details, see Online Appendix Table A1.

[7] Participants played a binary trust game in which we elicited responses for both roles; a risk attitude task (Gneezy and Potters, 1997); an honesty game (i.e. the coin flip task in Abeler et al. 2014); a series of 8 mini-dictator games (Charness and Rabin, 2002); and a series of attitudinal scales. No feedback was given to the subjects until the completion of the experiment on day 2. Upon the completion of the experiment, one of the games were randomly selected for payment.

[8] In the FEELE lab, all experiments require participants to be in the lab at the scheduled start of the session, rather than offering a time window for individuals to turn up. Invitation emails always specify the end time of the session; the end time is determined conservatively, in order to account for any unexpected events (e.g. computer network crash), so that the sessions do not run over their expected end time. We were deliberately conservative in setting the maximum session duration to be one hour given that technical problems with software may be harder to resolve in an online environment than in a physical lab. In that sense, the fact that the expected duration of the session was shorter than the maximum duration of 1 h was not out of the ordinary for our participants.

pants received no feedback about their payoffs from that session; the only message they received was that they would be contacted again via email regarding the day-two session.

A few hours after the end of the day-one session, we sent a reminder regarding the day-two session. This was the main experimental manipulation (see Experimental Materials Section A2). We manipulated the texts of the reminder email to implement our experimental treatment conditions. Across all treatments, the subjects were told to arrive on time for the day-two session as they would be interacting in real time with other participants, and being late would cause inconvenience. The email did not make reference to the recipient's own punctuality in day one. This email was then extended to include the follow sentences, depending on the treatment.

Social information - majority punctual:

We noted that the majority of participants with whom you will be matched in the second session logged on promptly at 9 am in the first session of the experiment.

Social information - majority late:

We noted that the majority of participants with whom you will be matched in the second session were late to log on by 9 am in the first session of the experiment.

Reward for punctuality:

(...) those who arrive on time (i.e. those who log in at most 1 min after 9am) will receive £1 in addition to their £3 show up fee for the second session.

Fine for tardiness:

(...) latecomers will forfeit £1 of their £3 show up fee for the second session if more than 1 min late.

Note that the fine or reward is equivalent to 33% of the £3 show-up bonus the subjects receive. The reminder emails to subjects in the control conditions included either only one or none of the sentences above. The link to the day-two session was included at the very end of the email, so all subjects were expected to read through the experimental manipulation.

Upon the completion of the tasks for the day-two session, the subjects were asked to complete an online survey. At the end of the second day, the subjects received a final email with their detailed payment information.

A discussion of our definition of punctuality and the size of the incentives is warranted at this point. The FEELE lab operates on a very strict policy with regards to punctuality, as late arrivals can compromise the normal functioning of an experimental session. As a result, it is at the discretion of the experimenter whether or not to allow participants who arrive more than a minute late to take part. Those who arrive late are not eligible for a show-up fee altogether. Three no-shows result in the exclusion from the ORSEE participant pool altogether. Consequently, we expected that the punctuality threshold was well understood by all participants. The fine/reward level in Study I was in fact less stringent than the existing lab policy. This was because there could be extenuating circumstances for late arrival (e.g. a bad internet connection).

We implemented a 3×3 between-subjects design where we varied both the social information (no info, majority late or majority punctual) and the monetary incentives (no incentives, reward or punishment). For expositional purposes, we will denote treatments using acronyms that combine the social information manipulations, N (No Info), ML (Majority Late) and MP (Majority Punctual) with incentive conditions: N (Nothing), F (Fine) and R (Reward). For example, our baseline condition, No Info-Nothing is denoted as N-N. Depending on the social information conditions, and conditional on day one behavior, we randomly assigned subjects to groups with different proportions of late and punctual people. To remain *truthful* to the social information nudge messages, the subjects who were punctual on day one were randomly but *not evenly* distributed across all treatments *by construction* (the same applies to those subjects who were late on day one). For instance, ML-x treatment would inevitably have higher proportion of participants who were late on day one of the study, while MP-x treatment would have higher proportion of participants who were punctual on day one. Furthermore, the ML-x conditions had on average larger sample sizes than MP-x conditions, due to the fact that fewer subjects (around 40%) attended the day one session on time. This particular construct of random treatment assignment has considerable implications on how we analyze the data as we will discuss in Section 3.3.

We recorded participants' exact login times, which is our outcome variable of interest. Our design combines ecological validity with experimental control typically associated with the laboratory, in that we observe behavior that is the product of actual trade-offs (as opposed to induced valuations), while retaining a great degree of control over the experimental environment.

3.2. Norms on punctuality in the sample

Is being punctual generally considered a desirable behavior in our sample? From the two post-experimental survey questions on punctuality, we find that the answer is yes: our subjects generally value punctuality highly in workplaces (8.7 out of 10) and believe that others are also highly likely to value punctuality (7.6 out of 10). Those answers did not differ much by treatment conditions.[9]

[9] More details can be found in Section A2 of the Online Appendix. Although there are occasional pairwise differences in means, a joint test of equality of means is marginally significant for own value of punctuality $F(8, 539) = 1.82$, $p = .070$ and non-significant for perceptions of others' valuation $F(8, 539) = 1.63$, $p = .115$.

Table 1
Descriptive statistics - Study I.

Treatments	% of Late Subjects		Change	N
	Day 1	Day 2		
N-N	67%	48%	-19%	87
N-F	63%	37%	-26%	83
N-R	62%	31%	-31%	81
ML-N	64%	40%	-24%	87
ML-F	73%	28%	-45%	125
ML-R	62%	21%	-41%	86
MP-N	48%	40%	-8%	62
MP-F	46%	19%	-27%	67
MP-R	49%	30%	-19%	61
Total	61%	33%	-28%	739

As a robustness check, we ran a new online survey study and elicited injuctive norms of punctuality with a separate sample from FEELE lab subjects pool using the instrument developed by Krupka and Weber (2013).[10] The data confirms our results: 95% of our sample believes punctuality is very socially appropriate or somewhat socially appropriate in workplace. We also elicited injunctive norms of punctuality in the FEELE lab using the same methodology: 91% of our sample believes punctuality is very socially appropriate or somewhat socially appropriate.

We also checked whether our definition of punctuality and the perceived appropriateness of penalizing tardiness was shared by our subject pool. We elicited the social appropriateness of punishing tardiness by £1 for each of xx-F conditions on a separate sample of participants using the Krupka and Weber (2013) method in the new survey study. In all scenarios around 90% of participants described fining £1 for tardiness either 'Very socially appropriate' or 'Somewhat socially appropriate' (see Table A12 in the Online Appendix for a breakdown).

Does the fact that the overwhelming majority of our subjects value punctuality (making punctuality an injunctive norm) lead to them being punctual to our sessions? The answer is no. In line with our experimental email reminders, we consider a subject as being late if her arrival time is strictly after 9:01am in either day. Table 1 summarizes descriptive statistics for all treatments; 61% of subjects were in fact late on the first day of the experiment. Those who were punctual on day one logged on average just under 7 min early; those who were late logged on average just under 14 min late.[11]

In short, we successfully generated an environment in which there is a desirable mode of behavior (i.e. punctuality) that our subjects value. However, the personal cost associated with engaging in that behavior is sufficiently high to prevent most subjects from doing it. This means that there is room for improvement using financial incentives, normative messages, or a combination of the two.

3.3. Data analysis strategy

A key feature of our experimental design is that we are able to identify different types of subjects in our sample based on their day one behavior and observe how different types respond to our treatment manipulations. Meanwhile, it is important to note that, by design, different treatments have various composition of late and punctual subjects. Consequently, we cannot compare the overall effect across different treatments directly or the comparisons would be biased. For example, it may appear that the social information treatments with majority late messages (MP-N, MP-F and MP-R) have overall smaller effects compared to other conditions. However, the social information treatments with majority punctual messages also have the lowest percentage of late subjects.

To account for the composition differences between treatments and to provide appropriate tests of our hypotheses, we condition our statistical analysis on subjects' day one behavior. Following the nomenclature used in our model, subjects who were late on day one are *unmotivated* types, and those who were punctual on day one are *motivated* types.[12] Subjects who were late on day one were about 30% more likely to be late again on day two than subjects who were punctual in the N-N condition. In short, our categorization captures a consistent mode of behavior.[13]

Given that our fine and reward conditions have the same directional prediction, irrespective of the interaction with social norm manipulations, for the sake of simplicity of presentation, we will pool the data from the fine and reward conditions. We report the disaggregated analysis in the Online Appendix Section A3.

[10] We thank an anonymous referee for this excellent suggestion. The new online survey study was conducted in December 2020. 56 subjects in total participated in this new survey study. None of them had previously participated in our Study I. The study lasted on average 10.28 min and the average payment was £5.

[11] See Fig. A2 in the Online Appendix for a distribution plot of the arrival times.

[12] This categorization strategy is coarse, yet effective: subjects' behavior in day one is highly predictive of their behavior on day two even after controlling for observable characteristics such as gender, age, and attitude towards punctuality (see Table 3).

[13] Of course, this categorization is potentially noisy, in that random shocks such as a poor internet connection may have stopped otherwise punctual subjects to be on time on day one. We report on a number of robustness checks on our measure of punctuality in the Online Appendix Section A5, and show that our results are qualitatively robust to different categorization of types.

Table 2
Estimated intervention effects on tardiness rate by subject type.

DV: $\mathbb{I}(\text{Late}_{i,t=2})$	Unmotivated Subjects	Motivated Subjects
Incentive	-0.170***	-0.117*
	(0.050)	(0.056)
ML	-0.105**	-0.055
	(0.050)	(0.058)
MP	-0.073	-0.048
	(0.059)	(0.058)
Incentive × ML	-0.017	0.007
	(0.100)	(0.117)
Incentive × MP	0.040	-0.097
	(0.123)	(0.118)
N	451	288

Note: The control group is NN-N. We report estimated discrete changes in the probability of arriving late as a function for each of the main dummy regressors. For ease of exposition, we report contrasts for the interactions, which should be interpreted as changes in probability as a function of changes in the ML/MP dummy keeping Incentive=1. They are not additive to the main effect. Standard errors in parentheses. $^{***},^{**},^{*}$: $p < .01$, $p < .05$, $p < .1$.

3.4. Main results - study I

We estimate the following model:

$$\mathbb{I}(\text{Late}_{i,t=2}) = \beta_0 + \beta_1 \text{ Incentive} + \beta_2 \text{ ML} + \beta_3 \text{ MP} + \beta_4 \text{ Incentive} \times \text{ML} + \beta_5 \text{ Incentive} \times \text{MP} + u_i \quad (6)$$

Our dependent variable, $\text{Late}_{i,t=2}$, equals one if participant i logged in late to the day two session and zero otherwise. Incentive equals one if participant i was assigned to a treatment that involved incentives (i.e. xx-F, or xx-R) and zero otherwise; ML equals one if participant i was assigned to any of the ML-x treatments and zero otherwise; MP equals one if participant i was assigned to any of the MP-x treatments and zero otherwise. We estimate this model for unmotivated and motivated subjects separately. We report marginal effects and contrasts based on Probit estimates of Eq. (6) in Table 2.

Result 1. Financial incentives and majority late social information have a significant and desirable impact on the behavior of unmotivated types.

We start by looking at the effect of our interventions on unmotivated types. We focus our analysis on interventions that have predicted effects: incentives, social information about majority punctual, and their interaction. We observe a statistically significant main effect from incentives. However, we do not find significant main effect from majority punctual social information nor any interaction effect between social information and incentives. It is worthwhile to note that the social information about bad behavior (majority late) has a significant effect on reducing tardiness among unmotivated subjects.

In particular, the estimated marginal effect of Incentive is negative and significant ($\chi^2(1) = 11.78$, $p < .001$). The unmotivated types were 17 percentage points less likely to arrive late on day two compared with the control, when financial incentives were present. Our data thus support Hypothesis 1.

The estimated coefficient on MP is negative but not statistically significant ($\chi^2(1) = 1.52$, $p = .218$). Our data rejects Hypothesis 2. Social information about good behavior did not seem to have a main effect on the behavior of the unmotivated types. Finally, our data rejects Hypothesis 4. The estimated coefficient on Incentive × MP is positive but not significant ($\chi^2(1) = 0.10$, $p = .747$).

Result 2. Only financial incentives have a weakly significant and desirable impact on the behavior of motivated types.

We now turn to the behavior of motivated subjects. Again, we focus our analysis on interventions that have predicted effects: social information about the majority being late. Overall, we do not find any significant main effect from social information or interaction effect. The estimated coefficient on ML is negative and not significant ($\chi^2(1) = 0.92$, $p = .337$). Our data therefore reject Hypothesis 3. The estimated coefficient on Incentive × ML is very close to zero and not significant ($\chi^2(1) = 0.00$, $p = .950$). As a result, our data reject Hypothesis 5.

It is worth highlighting two non-hypothesized effects. Firstly, the negative effect of ML on the likelihood of being tardy on day two among unmotivated subjects. Secondly the negative effect of incentives of motivated subjects. The latter effect has a straightforward explanation: incentives reinforce intrinsic motivation and prevent motivated subjects from 'drifting' into tardiness. The former is harder to explain without examining the effect of ML on beliefs. We will examine this effect next.

Table 3
Estimates of the effects of normative attitudes and beliefs on punctuality.

DV: $\mathbb{I}(\text{Late}_{i,t=2})$	(1)	(2)	(3)	(4)
$\text{Late}_{i,t=1}$	0.074**	0.076**	0.085**	0.087**
	(0.038)	(0.037)	(0.038)	(0.038)
InjNormPunctual	-0.113***	-0.119***	-0.109***	-0.114***
	(0.038)	(0.038)	(0.038)	(0.038)
BeliefsOthersPunctual	-0.298***	-0.300***	-0.305***	-0.305***
	(0.056)	(0.057)	(0.057)	(0.057)
Late × InjNormPunctual		-0.213***		-0.222***
		(0.077)		(0.077)
Late × BeliefsOthersPunctual		0.048		0.059
		(0.116)		(0.114)
Treatment Controls	No	No	Yes	Yes
N	548	548	548	548

Note: We report estimated discrete changes in the probability of arriving late on day two as a function for each of the main dummy regressors. For ease of exposition, we report contrasts for the interactions, which should be interpreted as changes in probability as a function of changes in the relevant dummy keeping Late=1. They are not additive to the main effect. Standard errors in parentheses. $^{***},^{**},^{*}: p < .01, p < .05, p < .1$.

3.5. Attitudinal data and the role of beliefs

In this section, following the model proposed in Section 2, we explore the motivations for subjects to arrive on time on day two. We do so by estimating the utility function we proposed, while allowing individuals to vary in their tastes and on observable characteristics. The main components of the utility function are: the direct and opportunity costs of arriving on time (c_i, B_i), the utility derived from abiding by the injunctive norm, $\phi_i S$, and the utility derived from following the crowd (social information), $\gamma_i N(\mu)$.

To measure the strength of the innate preference for punctuality, proxying $\phi_i S$, we use the subjects' answer to a post-experimental survey question: *"On a scale of 1 to 10, how much do you value being punctual in the workplace?"*. To infer the subjects' beliefs about others' punctuality (an approximation for μ) on day two, we asked our subjects to guess whether or not the majority of their group members arrived on time on day two, given their assigned treatment conditions. To make the guesses incentive-compatible, subjects received an additional £0.50 if their guess was correct. The final element of our utility function (c_i, B_i) are not directly observable or measurable; given that we have direct measures of all other elements, we will use timely arrival on day one as a proxy for high costs of arriving on time. 191 participants did not complete the post-experimental survey.

Table 3 summarizes marginal effect estimates on the probability of arriving late on day two based on a probit estimator. The dependent variable is a dummy variable indicating whether subject i was late on day two. Regression (1) has as regressors the dummy variable Late ($t = 1$), which equals to one if subject i was late on day one; the dummy variable *InjNormPunctual*, which equals to one if subject i's answer to the survey question on personal value for punctuality is above the sample median.[14] The binary variable *BeliefsOthersPunctual*, which equals to 1 if subject i made an incentivized guess that most of the other participants were punctual on day two.

We find that subjects are significantly more likely to be late if they: (i) have higher costs of arriving on time, as per the positive and significant coefficient on Late ($t = 1$); (ii) do not value punctuality innately (negative and significant coefficient on InjNormPunctual); (iii) believe that most of their group members would be late (negative and significant coefficient on BeliefsOthersPunctual).

As shown earlier, being late on day one is positively correlated with being late on day two. We find innate preference for punctuality has some explanatory power: subjects who stated a preference for punctuality above the sample median are 11% less likely to be late on day two vis-à-vis subjects who expressed below median preference for punctuality.[15] The variable capturing the beliefs on others has the largest effect: subjects are just under 30% less likely to be late if they believe that majority of their group members would be punctual. Adding controls for different treatments does not change the results, as shown in regression (3).

In regression (2), we added two interaction terms, Late × InjNormPunctual and Late × BeliefsOthersPunctual, to capture the differential responses from the motivated and unmotivated types. We find no significant differences between either type, although when we add treatment controls, we find a significant difference in terms of intrinsic preference for punctuality. Compared with the motivated types, unmotivated types are more likely to change their behaviors due to changes in the innate value of punctuality induced by different treatments. However, they are not significantly more likely to change their

[14] We note that this variable was not estimated in an incentive compatible way. It is therefore likely to be noisy. We do not have reason to believe that it is biased, since there are no incentives to lie about this belief.

[15] Using mean split yields qualitatively and quantitatively similar results.

Table 4
Estimates of intervention effects on beliefs by subject type.

DV: Belief about others' punctuality		
	Unmotivated Subjects	Motivated Subjects
Incentive	0.230***	0.066
	(0.074)	(0.050)
ML	-0.0005	-0.082
	(0.049)	(0.053)
MP	0.155***	0.061
	(0.047)	(0.050)
Incentive × ML	0.005	-0.079
	(0.145)	(0.111)
Incentive × MP	-0.233*	0.018
	(0.123)	(0.103)
N	307	241

Note: We report estimated discrete changes in the probability of arriving late on day two as a function for each of the main dummy regressors. For ease of exposition, we report contrasts for the interactions, which should be interpreted as changes in probability as a function of changes in the relevant dummy keeping Incentive=1. They are not additive to the main effect. Standard errors in parentheses. ***,** ,* : $p < .01$, $p < .05$, $p < .1$.

behavior as a function of their beliefs about others' punctuality. Regressions (3) and (4) replicate the analysis of regressions (1) and (2) by adding a series of dummy variables for each of the treatments to control for any unobserved heterogeneity driven by treatment assignment. The results are unchanged.

Observation 1.1: *Punctuality on day two is predicted by "intrinsic costs", by perceptions of what is socially appropriate, and by socially prevalent behavior. The unmotivated types are more sensitive than motivated types to changes in the injunctive norm.*

The strength and significance of our elicited beliefs variable suggests that beliefs about the punctuality of others may partially mediate the effect of our treatment intervention. To verify this, we regressed our belief measure about the punctuality of others on a dummy for incentives, the norm manipulations and their interactions. Table 4 summarizes the results.

We find that beliefs by unmotivated types about others' punctuality are sensitive to incentives and to the majority-punctual manipulation. When incentives are interacted with the majority punctual norm manipulation, the effect is actually negative and of a larger magnitude. The negative coefficient indicates that it maybe counter-productive when combining incentives and majority punctual nudge, although the coefficient is only marginally significant. In contrast, the beliefs by motivated types were unaffected by any intervention.

The observation that incentives have distinct effect on unmotivated and motivated types is not entirely surprising: unmotivated subjects believe incentives will lead others to become more punctual, while motivated subjects do not. Indeed, motivated types on average hold stronger beliefs about others' punctuality than unmotivated types (see Fig. A1 in the Online Appendix), and are less sensitive to treatment effects. The majority punctual manipulation works as predicted: unmotivated subjects likely revised their beliefs upwards upon being informed by the reminder email. More surprising is the lack of effect of the majority late manipulation among the motivated subjects; although the coefficient has the correct sign, it is not statistically significant.

Observation 1.2: *Incentives raise unmotivated subjects' belief about the punctuality of others, so does social information about other being punctual. However, the combination of incentives and social information is counterproductive. Motivated subjects' beliefs are unchanged by any intervention.*

3.5.1. Discussion

We have successfully generated an environment in which virtually all participants agreed upon a correct course of action (i.e. to arrive on time to an appointment), yet few people actually took that action. Furthermore, the data are broadly consistent with our theoretical model. We find strong evidence of different types amongst individuals: those who were intrinsically motivated to turn up on time and those who were not.

Injunctive norms around punctuality are strong predictors of punctuality: subjects with the stronger injunctive concerns were 11 percentage points less likely to be late than those with weaker injunctive norm concerns. Injunctive norms were particularly important for unmotivated types, as those who exhibited above median injunctive norm concerns were about 20 percentage points less likely to be late than unmotivated types who had below median injunctive norm concerns.

Beliefs about the punctuality of others were also strong predictors of behavior. Those who believed that most subjects would be punctual to the day two session were 30 percentage points less likely to be late than those who believe most subjects would be tardy. This is consistent with the laboratory evidence of Bicchieri and Xiao (2009), in a dictator game context.

In light of the above discussion on the importance of the beliefs of others' punctuality, we would expect social information based nudge to be particularly effective. Instead, we found that, with one exception, messages about the behavior

of other participants had a very small and non-significant effect on punctuality in both motivated and unmotivated types. One potential explanation is that our participants might not fully agree with our definition of tardiness. After all, penalizing tardiness on the basis of being one minute late could be perceived as harsh or inappropriate. However, data from the follow-up online survey study allow us to rule out this explanation (see Online Appendix Section A6 for more details).

One possible explanation for the under-performance of the information nudge may be observability. Bolton et al. (2020) find in the context of the dictator game that having a third player observe the giving decision leads to a decrease in dictator's pro-social behavior. All our information nudge treatments made it explicit to subjects that their behavior had been (and would be) monitored by a third party (the experimenter), while such observability was absent from the treatments without social information.

In comparison, incentives were very effective at promoting or retaining punctuality when done in isolation. Perhaps surprisingly, the interaction of information nudge and incentives yielded small and non-significant effects on behavior. Data on the determinants of beliefs about others' punctuality suggests that coordination motives may be one explanation: incentives boosted the beliefs of unmotivated types about the punctuality of others by 23 percentage points, as did the majority punctual information intervention (16 percentage points). Nonetheless, combining incentives with the majority punctual information only lowered the beliefs about others' punctuality relative to either intervention in isolation.

We theorized social information nudge as operating through a belief channel. The higher the proportion of people one believes is cooperating, the larger the psychological benefit from doing the same. The correlation between belief and psychological benefit introduces a coordination motive: if one believes enough people arrive on time, it is preferable to do the same. Our data suggest that such coordination considerations, albeit important, are not sufficient to promote good behavior.

One surprising result was that unmotivated subjects were responsive to negative social norms information. Given that the context of this intervention was attendance of an experimental session (in a lab where there are a strong rules in place around tardy attendance), we conjecture that the ML intervention may have prompted some tardy subjects to show up earlier out of fear of reputational consequences regarding future participation in FEELE lab studies. This is an unfortunate limitation given the FEELE lab policy.

Another limitation of Study I was the noisy behavior in the variable of interest. There may have been many circumstances outside the control of our participants that lead to variation in the outcome variable. This is reflected in the large change in behavior in the baseline condition where we introduced no norm manipulations or incentives: while 67% of subjects were late on day one, only 48% were late on day 2. This 19 percentage points change indicates noisy behavior and also makes identifying average treatment effects difficult. It is also the case that the externality that tardiness imposes to others is diffuse: it is difficult for subjects to compute explicitly the disutility they impose on others from the timing of their arrival to the experiment.

4. Study II

In the second experiment, we sought to understand whether the main results from study I carried over to a different context. In particular, we wanted to examine the predictions of our model in a decision-making environment where externalities were explicit, and where we could ensure that observability of actions was certain. We designed an environment that closely mimicked our first study: subjects made a binary decision over two days, and their decisions had consequences to a large number of other decision makers.

We applied Study I's experimental design to a large N, binary prisoners' dilemma. While the desirable action in Study I was to arrive on time, it was cooperation in Study II. Although distinct from punctuality, social dilemmas such as the prisoner's dilemma fit our general research question very well: they are environments in which there is a socially desirable action which comes at a cost to the individual. Bad behavior imposes a notably negative externality to society. Furthermore, there are many examples of social dilemma environments where social norm information and incentives are regularly used: tax compliance (Slemrod et al., 2001; Kleven et al., 2011; Hallsworth et al., 2017), public transport use (Gravert and Collentine, 2019) and cooperative employee behaviors such as organizational citizenship behavior (Deckop, Mangel, Cirka, 1999).

4.1. Experimental design

Study II was pre-registered and conducted between January and February 2019.[16] We recruited 1573 US-resident participants on MTurk, all of whom completed day one of the experiment; 260 subjects failed to return to undertake day 2. Of those that returned, we excluded 220 subjects as they did not respond appropriately to the attention checks and a further 34 because they failed to correctly answer the comprehension questions after three attempts (as proposed in the pre-registration plan). We placed a HIT on MTurk that advertised a study that would take place over two days. The HIT

[16] The experimental design pre-registration can be found here: https://osf.io/bmkw9/. We report the following deviation from the pregistration: while we planned to collect up to 2000 participants and use 1200 participants after the exclusion, we actually have 1059 participants in our sample after the exclusion. The reason for the deviation was due to the extremely high number of exclusions we experienced and the limited budget we had. It is important to note that the reasons for our deviation are independent from treatment allocation. Given the number of participants we recruited, they were randomly assigned to the different treatments.

Table 5
Descriptive statistics - study II.

Treatments	% Keep		Change	N
	Day 1	Day 2		
N-N	71%	68%	-3%	130
N-F	72%	40%	-32%	137
N-R	72%	32%	-40%	141
MK-N	71%	82%	11%	146
MK-F	70%	52%	-18%	139
MK-R	69%	41%	-28%	137
MI-N	46%	37%	-9%	81
MI-F	50%	23%	-27%	74
MI-R	49%	22%	-27%	74
Total	66%	48%	-18%	1059

informed that they would be paid a fee of 50c for each part of the study and may earn more depending on their choices and the choices of others. Subjects were also informed that they would only be paid if they completed both days. Fees for completion were automatically paid through the MTurk system once approved after the experiment was completed. Payments relating to choices from both days of the experiment were paid as bonuses through MTurk upon completion of the second day. The resulting sample consisted of 1059 subjects, 47% of whom reported being male. The average duration of the day one and day two sessions was 4 min and 2 min, 29 s, respectively. The average payment for both days including show up fees was $4.45.

In each of the sessions, subjects were endowed with $1. Their decision was whether or not to invest their endowment, or keep it. The instructions informed subjects that there were at least 35 people in the group and the marginal per capital return (MPCR) for the group account was 5% (the MPCR is strictly greater than $1/35 \approx 3\%$).

The experimental design was identical to Study I. There were in total nine treatments, which varied by financial incentive and social information. Similar to Study I, we denote treatments using acronyms that combine the social information manipulations, N (No Info), MK (Majority Keep) and MI (Majority Invest) with incentive conditions: N (Nothing), F (Fine) and R (Reward).

The timeline of the experiment is the same as that in Fig. 1, except that the invitation to participate was done on the same day of the experiment. As in Study I, assignment to treatment was done after the day one decision took place, so as to abide by the no deception principle. Different from Study I, experimental manipulation was done in the day two instructions for Study II instead of a reminder email. Depending on the treatment, the following sentences were added to the day two instructions.

Social information - majority invest:

We noted that the majority of participants with whom you are matched today invested their $1 yesterday.

Social information - majority keep:

We noted that the majority of participants with whom you are matched today kept their $1 yesterday.

Reward for investing in the group account:

Additionally, those who invest their $1 today will receive a further 33% in addition to their payment from the group account.

Fine for keeping the endowment:

Additionally, those who keep their $1 today will forfeit 33% of their payment from the group account.

It is worth to note an important difference how we implemented the incentives treatments in Study I and II. In study I, incentives (fine or reward) were a fixed amount of the show-up fee (£1 out of £3 or 33% of the show-up fee), while in study II they are a proportion of the payoff (33% of the payoff from the group account) from the prisoners' dilemma. There are two main reasons for the change. The first is that in study I there is no financial externality from punctuality, while in study II externality is inherent in the prisoners' dilemma – the more people invest, the higher the payoffs to everyone in the group. Since the treatments explicitly manipulate the distribution of contribution levels, it is plausible to assume that individual financial payoffs will differ substantially across treatments. Therefore, a fixed fine/reward would have different salience in different treatments. By making financial incentives proportional to cooperation returns, we make treatment comparisons cleaner. A second reason is procedural: even if we had wished to, we would not have been able to withdraw participation fees from MTurk subjects, as per the terms and conditions of the platform.

4.2. Norms on cooperation in the sample

We again start by checking whether cooperation in a group setting is considered a desirable behavior in our Amazon Mechanical Turk sample. To this effect, we asked two post-experimental survey questions about whether subjects value cooperation, and whether they believe other subjects value cooperation. We obtain a positive answer in both cases, with a mean response rate of 7.4 out of 10 for the first question and 6.6 out of 10 for the second. However, the majority of subjects were uncooperative despite them valuing cooperation. Table 5 summarizes the descriptive statistics for all treatments: 66% of the participants kept their endowment on day one, and 48% of the participants kept their endowment on day two of the

Table 6
Estimated intervention effects by subject type.

DV: $\mathbb{I}(\text{Kept}_{i,t=2})$	Unmotivated Subjects	Motivated Subjects
Incentive	-0.353***	-0.197***
	(0.031)	(0.040)
MK	0.099***	0.171***
	(0.037)	(0.050)
MI	-0.117**	-0.026
	(0.053)	(0.048)
Incentive × MK	0.057	-0.303***
	(0.061)	(0.095)
Incentive × MI	0.056	0.238***
	(0.100)	(0.074)
N	696	363

Note: We report estimated discrete changes in the probability of keeping one's endowment on day two as a function for each of the main dummy regressors. For ease of exposition, we report contrasts for the interactions, which should be interpreted as changes in probability as a function of changes in the relevant dummy keeping Incentive=1. They are not additive to the main effect. Standard errors in parentheses. $^{***},^{**},^{*}$: $p < .01$, $p < .05$, $p < .1$.

experiment. That is, we again generated an environment in which there is a mode of behavior that is widely regarded as desirable, yet only a minority choose to engage in.

Following the nomenclature in our model, subjects who kept their endowment on day one are *unmotivated* types, and those who invested their endowment on day one are *motivated* types. Our type categorization maps on to the usual typology in cooperation games (Fischbacher et al., 2001). Motivated types correspond to unconditional cooperators, and unmotivated types map on to defectors or conditional cooperators. While we understand our typology is unusual in the context of cooperation games, our objective is not to understand how our experimental manipulations affect conditional cooperators and defectors differently; therefore, for the sake of consistency with Study I, we will retain the same nomenclature as the rest of the paper.

4.3. Main results - study II

We now examine the effect of the different treatment manipulation on cooperation on day two, conditional on behavior on day one. Table 5 displays the proportion of subjects who chose to keep their endowment on day one and two. We note that behavior is much more stable over time in the baseline N-N treatment in study II than in study I: the proportion of subjects who kept their endowment only changed by 3 percentage points over the two days (in Study I, the changes from day one to two were 19 percentage points. We again note that the sample size in the MI treatments is smaller, because two thirds of subjects decided to keep their endowment.

Our analysis will rely on a series of regressions of the form:

$$\mathbb{I}(\text{Kept}_{i,t=2}) = \beta_0 + \beta_1\ \text{Incentive} + \beta_2\ \text{ML} + \beta_3\ \text{MP} + \beta_4\ \text{Incentive} \times \text{ML} + \tag{7}$$

$$+\beta_5\ \text{Incentive} \times \text{MP} + u_i \tag{8}$$

Our dependent variable, $\text{Kept}_{i,t=2}$, equals one if participant i kept his \$1 endowment on day two and zero otherwise. Incentive equals one if participant i was assigned to a treatment that involved incentives (i.e. xx-F, or xx-R) and zero otherwise; MK equals one if participant i was assigned to any of the MK treatments and zero otherwise; MI equals one if participant i was assigned to any of the MI treatments and zero otherwise. We estimate this model for the unmotivated and motivated sub-samples separately. As in study I, we report estimated marginal effects in Table 6.

Result 3. Financial incentives and social information have a significant and desirable effect on the behavior of the unmotivated types.

We start by looking at the effect of our interventions on unmotivated types. We focus our analysis on interventions that have predicted effects: incentives, social information about majority invest, and their interaction. We observe a statistically significant main effect from financial incentives ($\chi^2(1) = 129.91$, $p < .001$): compared to the control condition, subjects are 35 percentage points more likely to invest when financial incentives are present. This result supports Hypothesis 1. We also find a smaller yet significant effect on social information that the majority invested their endowment ($\chi^2(1) = 4.87$, $p = .027$): compared to the control condition, people are 11 percentage points more likely to invest when social information on good behavior is present. Our data therefore supports Hypothesis 2. We do not find any significant interaction effect between positive social information and incentives ($\chi^2(1) = 0.32$, $p = .573$). Consequently, our data reject Hypothesis 4.

Table 7
Estimates of behavioral, attitudinal and belief determinants of cooperation.

DV: $\text{Kept}_{i,t=2}$	(1)	(2)	(3)	(4)
Kept($t=1$)	0.319***	0.323***	0.341***	0.346***
	(0.029)	(0.030)	(0.028)	(0.030)
InjNormInvest	-0.132***	-0.134***	-0.120***	-0.122***
	(0.025)	(0.025)	(0.024)	(0.024)
BeliefsOthersInvest	-0.507***	0.507***	-0.454***	-0.454***
	(0.027)	(0.027)	(0.032)	(0.032)
Kept × InjNormInvest		-0.116**		-0.122**
		(0.055)		(0.054)
Kept × BeliefsOthersInvest		-0.142**		-0.152**
		(0.062)		(0.061)
Treatment controls	No	No	Yes	Yes
Observations	1059	1059	1059	1059

Note: We report estimated discrete changes in the probability of keeping one's endowment on day two as a function for each of the main dummy regressors. For ease of exposition, we report contrasts for the interactions, which should be interpreted as changes in probability as a function of changes in the relevant dummy keeping Kept=1. They are not additive to the main effect. Standard errors in parentheses. ***,**,*: $p < .01$, $p < .05$, $p < .1$.

Result 4. Social information (majority kept) has a significant but undesirable effect on the behavior of the motivated types. However, social information (majority kept) and incentives have a significant and desirable interaction effect on the behavior of the motivated types.

We now turn to the behavior of motivated types. Again, we focus our analysis on interventions that have predicted effects: social information about majority keeping, and its interaction with incentives. Table 6 shows that the estimated coefficient on MK is positive and significant ($\chi^2(1) = 24.03$, $p < .001$). This result supports Hypothesis 3: introducing social information that majority kept their endowment significantly reduced the likelihood of cooperation among the motivated types. The coefficient on *Incentive* × *MK* is negative and significant ($\chi^2(1) = 10.20$, $p = .001$) Our data therefore supports Hypothesis 5.

There are three non-hypothesized results that merit a brief mention. In the unmotivated subject sub-sample, it is noteworthy that we estimated a positive and significant coefficient on MK. The negative coefficient indicates that introducing negative social information increases the likelihood of non-cooperation among unmotivated subjects ($\chi^2(1) = 7.32$, $p = .007$). It is conceivable that unmotivated subjects who would otherwise be cooperative on day two are less likely to be so when exposed to negative social information. In the motivated subject sub-sample, we find a negative and significant coefficient on *Incentive* ($\chi^2(1) = 24.03$, $p < .001$). This result suggests that incentives may help prevent those motivated subjects who would 'lapse' into non-cooperation (for reasons outside the scope of the model) from doing so. Finally, the coefficient on Incentive × MI is positive and significant. The positive coefficient implies that there may be a crowding out effect of intrinsic motivation when incentives are combined with social information nudge. Our finding is consistent with the crowding out effects that have been widely documented by the literature (see, e.g, Gneezy et al., 2011, 2018, List et al., 2018).

4.4. Attitudinal data

We use attitudinal data collected during the experiment in Study II to better understand the mechanisms behind the reported results above—importantly, we did not have any attrition when collecting the post-experiment survey data in Study II. We estimate the model proposed in Eq. (1). Table 7 reports the regression on the decision to keep one's endowment on day two. Consistent with Study I, a similar set of regressors were included. *Keep on D1* is a dummy variable, which equals to 1 if participant *i* kept her endowment on day one and 0 otherwise. This variable approximates the innate preference for punctuality for motivated and unmotivated types. *InjNormInvest* is another dummy variable. It equals to one if participant *i*'s answer to the survey question *"On a scale from 1 to 10, how much do you value being cooperative in the context of a group task?"* is above the median value of the population and 0 otherwise. *BeliefsOtherInvest* is alos a binary variable. It equals to one if participant *i* made an incentivized guess that most of the other participants in her group chose to invest their $1 in the group account on day two and 0 otherwise. Table 7 summarizes the estimation results.

Regression (1) estimates show that a participant is significantly more likely to keep the endowment if s/he belongs to the unmotivated type: the coefficient on Keep on D1 is positive and highly significant ($\chi^2(1) = 123.21$, $p < .001$), A participant is less likely to keep the endowment if s/he values group cooperation and investment as an injunctive norm: the coefficient on InjNormInvest is negative and highly significant ($\chi^2(1) = 28.39$, $p < .001$). A participant is also less likely to keep if s/he believes that majority of the group members would keep their endowment: the coefficient on BeliefsOtherInvest is negative and highly significant ($\chi^2(1) = 360.99$, $p < .001$). To understand the differential responses from the motivated and unmotivated types, in regression (2) we added two interaction terms: Keep × InjNormInvest and Keep × BeliefsOthersInvest. The coefficient on Keep × InjNormInvest is negative and significant ($\chi^2(1) = 4.52$, $p = .034$). The coefficient on

Table 8
Estimated intervention effects on beliefs by subject type.

DV: Belief about others' cooperativeness		
	(Unmotivated Subjects)	(Motivated Subjects)
Incentive	0.338***	0.218***
	(0.032)	(0.043)
MK	-0.121***	-0.324***
	(0.037)	(0.059)
MI	0.392***	0.184***
	(0.048)	(0.054)
Incentive × MK	-0.025	0.204**
	(0.064)	(0.096)
Incentive × MI	-0.227**	-0.282***
	(0.090)	(0.074)
N	696	363

Note: We report estimated discrete changes in the probability of keeping one's endowment on day two as a function for each of the main dummy regressors. For ease of exposition, we report contrasts for the interactions, which should be interpreted as changes in probability as a function of changes in the relevant dummy keeping Incentive=1. They are not additive to the main effect. Standard errors in parentheses. ***, **, * : $p < .01$, $p < .05$, $p < .1$.

Keep × BeliefsOthersInvest is also negative and significant ($\chi^2(1) = 5.24$, $p = .022$). In other words, the role of injunctive norms and beliefs is different between sub-samples: unmotivated types are more sensitive to both injunctive norms and beliefs about others' behavior than motivated ones. Regressions (3) and (4) replicate the analysis of regressions (1) and (2) by adding a series of dummy variables for each of the treatments to control for any unobserved heterogeneity driven by treatment assignment. The results remain unchanged.

Observation 2.1: *Investment in the group account on day two is predicted by "intrinsic costs", by perceptions of what is socially appropriate, and by socially prevalent behavior.*

4.5. The role of beliefs

We again explore the role of beliefs and how they are affected by the experimental manipulations. We regressed our belief measure about the cooperativeness of others on a dummy for incentives, the norm manipulations and their interactions. Table 8 summarizes the results. Marginal effects and interaction contrasts are reported.

Among unmotivated subjects, the positive coefficient on Incentive shows that introducing incentives to cooperate lead to an increase in beliefs about others' cooperativeness ($\chi^2(1) = 110.28$, $p < .001$); information nudge about majority invested achieves the same directional result ($\chi^2(1) = 66.04$, $p < .001$). Interestingly, the interaction between incentive and majority invested information is negative and significant ($\chi^2(1) = 6.41$, $p = .011$). This negative coefficient suggests that subjects in our experiment might be anticipating the crowding out effect on intrinsic motivation from financial incentives. Majority kept information exhibits a significant effect on beliefs in isolation ($\chi^2(1) = 10.67$, $p = .001$) but not when interacted with incentives.

Motivated subjects' beliefs about the cooperativeness of others were highly sensitive to incentives ($\chi^2(1) = 26.23$, $p < .001$) and both types of social information messages (majority invested: ($\chi^2(1) = 11.75$, $p < .001$; majority kept: $\chi^2(1) = 30.55$, $p < .001$). The interaction of information and incentives shift beliefs in the hypothesized direction in the case of majority invested information nudge ($\chi^2(1) = 14.50$, $p < .001$). In the case of majority kept information nudge, the data suggests the nudge effect dominates ($\chi^2(1) = 4.52$, $p = .034$).

4.6. Discussion

Study II confirms that behavior in the game is driven by types, injunctive norms around cooperation, as well as beliefs about others' behavior. It supports the finding from Study I that incentives are effective promoters of good behavior for both motivated and unmotivated types. However, in a social dilemma environment, the information nudges successfully induced behavioral change on both types. Negative social information (majority kept) led to an increase of free-riding behavior among both types of players. In contrast, positive information interventions (majority invested) were only effective on unmotivated types. Interactions of information nudges and incentives worked in motivated subjects: as predicted when interacted with negative information, and positively when interacted with positive information.

The greater effectiveness of information based nudges in this domain relative to punctuality is not surprising: beliefs about the distribution of types one is playing with have clear payoff consequences to subjects in an N-player prisoners' dilemma. This finding entails that some unmotivated types may be conditional cooperators: moving beliefs through information nudge campaigns can be effective for these types of individuals in environments where externalities are financially salient. As such the norm effect on beliefs includes inherently an income effect which was absent in the punctuality envi-

ronment: a conditional cooperator will want to cooperate more, the more cooperators s/he believes there to be in the pool of participants.

The belief channel accounts for the change in behavior as a result of interactions of social norm interventions and incentives among motivated types. It does not account for the null effect on behavior when positive information (majority invested) are combined with incentives.

5. Conclusion

We present results from two experiments through which we examine how financial incentives interact with social information to impact behavior. Our first study is a field experiment that combines a high degree of external validity with the experimental control typically associated with the laboratory — effectively a "field-in-the-lab" experiment. Our outcome variable of interest is whether or not our participants arrived on time to a sequence of experimental sessions that started at 9am on consecutive days. Importantly, our participants were unaware that measuring timely attendance was the object of the study. Punctuality is a mode of behavior which is universally understood and valued across societies. Our second study is a lab experiment with non-student sample, in which participants participated two large N, one-shot prisoners' dilemmas with at least 35 other subjects. Large N prisoners' dilemmas capture the fundamental trade-offs in many relevant policy environments, such as tax compliance, littering, speeding among many others. The loss of realism is compensated by a greater control over the measurement of the negative externality caused by bad behavior.

In both experiments, virtually all of participants valued good behavior (i.e. punctuality and cooperation, respectively), and believed others also valued it. However, most of our subjects failed to engage in such socially desirable action. Moreover, subjects displayed persistent behavior: in the absence of any intervention, participants were more likely to repeat their day-one behavior on day two. Both experiments exhibit a good baseline environment on which we base to test the effectiveness of financial incentives, social information and their interaction on behavior.

In both experiments, we categorized the subjects into unmotivated and motivated types based on their day one behavior. Not only did both types of subjects differ in their likelihood of behaving appropriately, but they also differed in their response to treatments. These differences were not only behavioral in nature, but also attitudinal. In the punctuality study, motivated types were unresponsive to most interventions. Importantly, they were unlikely to become tardy even when informed that the majority of their counterparts was late on the first day. These subjects had stronger beliefs about the punctuality of others, and those beliefs were less likely to be affected by incentives and social information than the beliefs of subjects who were not punctual on day one. In the social dilemma, motivated types were more responsive to interactions of incentives and norm interventions. Depending on the distribution of types and context, the aggregate effect of a given policy might be markedly different. This is the first substantial contribution of the present paper: we document systematic differences in average treatment effect on the basis of compliance types.

Further, our findings highlight the importance of understanding the potential heterogeneity of responses to treatment effects when designing public policy interventions. Hallsworth et al. (2017) find important differences in how tax debtors respond to norm interventions as a function of the size of their debt. Hauser et al. (2018) use machine learning methods to predict the likelihood of disclosure of past employment income by unemployment benefit applicants. They find those with a high risk profile are more likely to respond positively to behavioral interventions exhorting honest reporting of income, but some lower-risk cohorts may respond negatively. While our experimental design overcame this information asymmetry by design, judicious use of machine learning techniques (Athey and Imbens, 2015) to predict responsiveness to treatment based on observables could be a useful step in this direction.

There were important differences in the effectiveness of information nudges: social information nudges alone were ineffective at changing punctuality, while they were very effective in the prisoners' dilemma. The focus theory of normative conduct in social psychology (Cialdini et al., 1990; Bolton et al., 2020) argues that social norms motivate behavior when they are made salient to decision makers. We make a complementary observation: the externality generated by the behavior targeted by the norm information intervention also needs to be salient (either financially or in utility terms). If not, the norm information salience alone may not be sufficient. Our results suggest that behaviors whose social externalities are diffuse (in the sense that they might be difficult to compute or to observe) might be difficult to shift using norm information nudges, so incentives should be used instead. In contrast, norm information nudges are likely to be extremely effective in environments where externalities are easy to compute, such as tax compliance Hallsworth et al. (2017) and Bott et al. (2020). This is the second contribution of our paper: policy makers may wish to make explicit the social and/or financial consequences of bad behavior to individuals in order to maximize the effectiveness of any norm-based intervention.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.euroecorev.2021.103711.

References

Akerlof, G.A., 1980. A theory of social custom, of which unemployment may be one consequence. Q. J. Econ. 94 (4), 749–775.
Allcott, H., 2011. Social norms and energy conservation. J. Publ. Econ. 95 (9–10), 1082–1095.

Allcott, H., Kessler, J.B., 2019. The welfare effects of nudges: a case study of energy use social comparisons. Am. Econ. J. Appl. Econ. 11 (1), 236–276.
Athey, S., Imbens, G.W., 2015. Machine learning methods for estimating heterogeneous causal effects. Stat 1050 (5), 1–26.
Bénabou, R., Tirole, J., 2006. Incentives and prosocial behavior. Am. Econ. Rev. 96 (5), 1652–1678.
Bicchieri, C., Dimant, E., 2019. Nudging with care: the risks and benefits of social information. In: Public Choice, 1–22.
Bernheim, B.D., 1994. A theory of conformity. J. Polit. Econ. 102 (5), 841–877.
Bicchieri, C., Xiao, E., 2009. Do the right thing: but only if others do so. J. Behav. Decis. Mak. 22 (2), 191–208.
Blumenthal, M., Christian, C., Slemrod, J., 2001. Do normative appeals affect tax compliance? evidence from a controlled experiment in minnesota. Natl. Tax J. 54 (1).
Bolton, G., Dimant, E., Schmidt, U., 2020. When a nudge backfires: Combining (IM) plausible deniability with social and economic incentives to promote behavioral change. In: CESifo Working Paper No. 8070.
Bott, K.M., Cappelen, A.W., Sorensen, E., Tungodden, B., 2020. You've got mail: Aa randomized field experiment on tax evasion. Manag. Sci. 66 (7), 2801–2819.
Bursztyn, L., González, A.L., Yanagizawa-Drott, D., 2020. Misperceived social norms: women working outside the home in saudi arabia. Am. Econ. Rev. 110 (10), 2997–3029.
Burtch, G., Hong, Y., Bapna, R., Griskevicius, V., 2018. Stimulating online reviews by combining financial incentives and social norms. Manag. Sci. 64 (5), 2065–2082.
Charness, G, Rabin, M, 2002. Understanding social preferences with simple tests. Q. J. Econ. 117 (3), 817–869.
Cialdini, R.B., Reno, R.R., Kallgren, C.A., 1990. A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. J. Pers. Soc. Psychol. 58 (6), 1015–1026.
Cialdini, R.B., Trost, M.R., 1998. Social influence: Social norms, conformity and compliance. In: Gilbert, D.T., Fiske, S.T., Lindzey, G. (Eds.), The Handbook of Social Psychology. Boston: McGraw-Hill, pp. 151–192.
Deckop, J.R., Mangel, R., Cirka, C.C., 1999. Getting more than you pay for: organizational citizenship behavior and pay-for-performance plans. Acad. Manag. J. 42 (4), 420–428.
Dimant, E., van Kleef, G.A., Shalvi, S., 2020. Requiem for a nudge: framing effects in nudging honesty. J. Econ. Behav. Organ. 172, 247–266.
Galbiati, R., Schlag, K.H., Van Der Weele, J.J., 2013. Sanctions that signal: an experiment. J. Econ. Behav. Organ. 94, 34–51.
Gneezy, U., Meier, S., Rey-Biel, P., 2011. When and why incentives (Do not) work to modify behavior. J. Econ. Perspect. 25 (4), 191–210.
Gneezy, U., Potters, J., 1997. An experiment on risk taking and evaluation periods. Q. J. Econ. 112 (2).
Gneezy, U., Rustichini, A., 2000a. A fine is a price. J. Legal. Stud. 29 (1), 1–17.
Gneezy, U., Rustichini, A., 2000b. Pay enough or don't pay at all. Q. J. Econ. 115 (3), 791–810.
Gravert, C., Collentine, L., 2019. When nudges aren't enough: Incentives and habit formation in public transport usage. Mimeo.
Greiner, B., 2015. Subject pool recruitment procedures: organizing experiments with ORSEE. J. Econ. Sci. Assoc. 1 (1), 114–125.
Hallsworth, M., Chadborn, T., Sallis, A., Sanders, M., Berry, M., Greaves, F., Clements, L., Davies, S.C., 2016. Provision of social norm feedback to high prescribers of antibiotics in general practice: a pragmatic national randomised controlled trial. The Lancet 387 (10029), 1743–1752.
Hallsworth, M., List, J.A., Metcalfe, R.D., Vlaev, I., 2017. The behavioralist as tax collector: using natural field experiments to enhance tax compliance. J. Publ. Econ. 148, 14–31.
Hauser, O., Greene, M., De Celles, K., Norton, M., Gino, F., 2019. Minority report: A big data approach to organizational attempts at deterring unethical behavior. In: Proceedings of the Academy of Management Global. Vol. Surrey, No. 2018
Health, T. f. A., 2007. F as in fat: How obesity policies are failing in america. Available at http://healthyamericans.org/reports/obesity2007/.
Kleven, H. J., Knudsen, M. K., Kreiner, C. T., Pedersen, S., Saez, E., 2011. Unwilling or unable to cheat? Evidence from a tax audit experiment in Denmark.. Econometrica 79 (3), 651–692.
Kölle, F., Lane, T., Nosenzo, D., Starmer, C., 2020. Promoting voter registration: the effects of low-cost interventions on behaviour and norms. Behav. Publ. Pol. 4 (1), 26–49.
Krupka, E.L., Weber, R.A., 2013. Identifying social norms using coordination games: why does dictator game sharing vary? J. Eur. Econ. Assoc. 11 (3), 495–524.
Kullgren, J. T., Harkins, K. A., Bellamy, S. L., Gonzales, A., Tao, Y., Zhu, J., Volpp, K.G., Asch, D.A., Heisler, M., Karlawish, J., 2014. A mixed-methods randomized controlled trial of financial incentives and peer networks to promote walking among older adults. Healthy Aging 41 (15), 43–50.
List, J.A., Livingston, J.A., Neckermann, S., 2018. Do financial incentives crowd out intrinsic motivation to perform on standardized tests? Econ. Edu. Rev. 66, 125–136.
Myles, G.D., Naylor, R.A., 1996. A model of tax evasion with group conformity and social customs. Eur. J. Polit. Econ. 12 (1), 49–66.
Nolan, J.M., Schultz, P.W., Cialdini, R.B., Goldstein, N.J., Griskevicius, V., 2008. Normative social influence is underdetected. Pers. Soc. Psychol. B. 34 (7), 913–923.
Pellerano, J. A., Price, M.K., Puller, S. L., Sanchez, G. E., 2017. Do extrinsic incentives undermine social norms? evidence from a field experiment in energy conservation. Environ. Res. Eco. 67, 413–428.
Richter, I., Thogersen, J., Klöckner, C.A., 2018. A social norms intervention going wrong: boomerang effects from descriptive norms information. Sustainability 10, 2848.
Silva, A., John, P., 2017. Social norms don't always work: an experiment to encourage more efficient fees collection for students. PLoS: ONE 12 (5), p.e0177354.
Slemrod, J., 2007. Cheating ourselves: the economics of tax evasion. J. Econ. Perspect. 21 (1), 25–48.
Slemrod, J., Blumenthal, M., Christian, C., 2001. Taxpayer response to an increased probability of audit: evidence from a controlled experiment in Minnesota.. J. Public Econ. 79 (3), 455–483.
Thaler, R.H., Sunstein, C.R., 2008. Nudge: Improving decisions about health, wealth, and happiness. Yale University Press, New Haven, CT.
Thorndike, A., Riis, J., Levy, D.E., 2016. Social norms and financial incentives to promote employees' healthy food choices: A randomized controlled trial. Preventive Medicine 86, 12–18.
Van Der Weele, J., 2012. The signaling power of sanctions in social dilemmas. J. Law Econ. Organ. 28 (1), 103–126.