Social Norms with Private Values: Theory and Experiments[1]

Giovanna d'Adda, Martin Dufwenberg, Francesco Passarelli & Guido Tabellini[2]

July 27th, 2020

Abstract

We propose a simple theory of social norms that models the distinct influence on behavior of personal values, normative expectations and empirical expectations. The first and second moments of the distribution of normative expectations affect the strength of social norms' pull on behavior. We test the empirical predictions of the model through an experiment based on a variant of the dictator game. Consistent with the theory, we find that normative expectations influence generosity and that higher dispersion of such expectations leads to more variation in giving behavior.

Keywords: Social norms, partial norms, normative expectations, consensus, experiment.
JEL codes: C91, D91

[1] We are grateful to two referees for helpful comments, to PRIN grant 2015PJL3XP and to MIUR grant "Dipartimenti di Eccellenza 2018-2022" for financial support, and to Giacomo Lemoli, Francesca Miserocchi, and Tommaso Cohen for excellent research assistance.

[2] Respectively affiliated with Università Statale di Milano, EIEE, LdA and Fondazione Pesenti; University of Arizona, University of Gothenburg, and CESIfo; Università di Torino, Baffi Center, and CESIfo; Università Bocconi – Dept. of Economics and IGIER, CEPR, CESIfo. Emails, respectively: giovanna.dadda@unimi.it; martind@eller.arizona.edu; francesco.passarelli@unito.it; guido.tabellini@unibocconi.it.

1. Introduction

Many scholars argue that social norms powerfully influence behavior. Fehr & Schurtenberger (2018) present an engaging recent discussion. Following the pioneering work of Bicchieri (2006), a social norm is conceptualized as a commonly known standard of behavior that is based on widely shared views of how individual group members ought to behave in a given situation (see also Elster 1989). The standard of behavior established by a social norm is based on the expectation of how others actually behave (what Bicchieri 2006 calls empirical expectations) and how they think one ought to behave (normative expectations). Moreover, this standard of behavior is commonly known, shared by a *large enough* number of individuals, and followed conditional on both empirical and normative expectations. By entailing an aspect of conformity with the normative standards of others, social norms are distinct from individual values or preferences, which instead do not logically require to conform to the expectations of others; and from conventions and fashions, which miss any normative aspect (Bicchieri 2006).

This definition of a social norm leaves open the question of how strong others' compliance with empirical and normative expectations must be, for a social norm to exist. Bicchieri (2017) explicitly discusses this issue, acknowledging both that there may be a substantial range of behaviors deemed appropriate under a certain norm, and that the threshold of compliance necessary for a norm to exist may vary across individuals and behaviors. These elements of uncertainty in the content and scope of a specific norm may reduce its pull on behavior and diminish norm compliance.

In this paper we study theoretically and empirically how the normative and positive consensus on the content of a norm influence its strength. We formalize an approach to social norms that builds on the important work of Bicchieri (2006) on normative and empirical expectations, and we submit it to an experimental test. Our model formalizes two aspects of leading definitions of social norms. First, we propose a definition of social norms that makes explicit the distinction between empirical and normative expectations. Second, given the crucial role of the normative aspect in the definition of social norms, we formalize the notion of the level of uncertainty around a social norm in terms of properties of the distribution of normative expectations, and study its implications for individual compliance with the norm. We do not offer a generally applicable model, but rather focus on a version of the dictator game.

Our approach mirrors the one by Bicchieri (2006) in identifying three components which, besides his material incentives, affect an individual's behavior: (i) his subjective values (personal normative beliefs); (ii) the perceived values of others (normative expectations); (iii) the expected behavior of others (empirical expectations). Subjective values are necessary in our framework to derive the (normative and empirical) expectations of others. Individuals are allowed to have different subjective values, and this disagreement is common knowledge. Moreover, a stronger disagreement reduces the traction of the perceived values of others. Intuitively, if there is widespread disagreement on what is “the right thing to do”, then I care less about how others expect me to behave. This is how we capture the idea in Bicchieri (2017) that the influence of a social norm on behavior is stronger if it is widely shared.

The model yields two predictions. First, individual behavior moves in the direction of the average perceived values of others. Second, the perceived normative disagreement also matters, in subtle ways: a stronger perceived disagreement increases the dispersion in individual actions, but its effect on individual behavior also depends on the contrast between subjective values and the perceived values of others, with possibly counterintuitive effects. In our experiment we test these predictions. We induce an exogenous change in individual perceptions of the normative standards of others. We then ask how individual behavior is influenced: (i) by the average perceived normative standards of others; (ii) by perceived disagreement in such normative standards.

We test the model’s predictions experimentally in a variant of the dictator game, where the amount given to the other party is doubled by the experimenter. This doubling of the amount given adds some ambiguity over the contents of the social norm, relative to a simple dictator game where the equal split is a natural focal point. We change perceptions over the normative standards of others through an informational treatment showing participants different distributions of answers to the question on the socially appropriate behavior of the dictator in such a game.[3] We compare a default distribution of answers with two treatment distributions. First, a distribution that differs from the default in that it has a lower average amount given as the right thing to do, but the same variance – we call this the Low Average treatment. Second, a distribution that differs from the default in that it has a higher variance of answers, but has the same mean as the default distribution – we call this the High Variance treatment. We conduct two variants of the experiment: one where participants play the dictator game immediately after the information treatment; and one where we elicit participants’ values, normative and empirical expectations before asking them to play the game.

We obtain two results. First, individuals who received the Low Average informational treatment give less as dictators, compared to individuals exposed to the default distribution; while individuals who received the High Variance treatment on average do not donate a different amount than the default group, but exhibit a higher variance in the amount given, as a group. Second, the informational treatments affect beliefs in the expected direction, although the effect is strongest on

[3] These answers were drawn from a pilot study of the same experiment.

the beliefs over the normative standards of others. Namely, individuals exposed to the Low Average treatment perceive others as having the normative expectation of a lower amount given, while those exposed to the High Variance treatment have a more dispersed perceived distribution of the normative expectation of others. Moreover, these beliefs are correlated with actions in the expected manner. Overall, therefore, these results confirm that social norms influence actions through individual perceptions of the normative standards of others, and that such beliefs react to available information. Moreover, the consensus around the social norm, measured by the variance of perceived normative beliefs, also matters. A more partial norm (i.e., one on which consensus is weaker) is associated with more dispersion in individual actions, suggesting that the social norm is less influential, consistent also with the ideas in Bicchieri (2017). However, less consensus around the contents of a social norm need not imply more selfish behavior, since a weaker social norm also gives more influence to personal values.

We contribute to the existing literature in a number of ways. We make two main contributions to the theoretical literature on social norms. Several scholars (e.g. Lewis 1969, Taylor 1987, Sugden 1989, Landa 2006, Bicchieri 2006, López-Pérez 2008, Binmore 2010, Gintis 2010) offer general ways of thinking about social norms in a game-theoretic context.[4] Some of these theoretical frameworks, notably Bicchieri (2006) and Cialdini et al. (1990), conceptualize the distinction between normative and empirical expectations and discuss the interaction between them, both when they are aligned and when they are in contrast. However, these models do not include a formalization of these two types of expectations. Our first contribution to the theoretical literature on social norms is thus modeling explicitly these different influences on norm compliant behavior.

Second, we provide a formalization of the notion of partial consensus around the social norm and study its implications. Bicchieri (2017) discusses this issue, arguing that ambiguity around the content of a social norm is likely to lead to lower norm compliance and to norm manipulation, whereby individuals select the interpretation of the norm that yields highest private returns. In a related discussion, Miller & Prentice (2016) argue that individuals infer the strength of a social norm from the extent of others' norm-compliant behavior. In our setting, instead, normative disagreement does not necessarily lead to more selfish behavior. The reason is that more normative disagreement

[4] Lewis (1969), Taylor (1987), Sugden (1989) and Landa (2006) propose ways of thinking about social norms in terms of rationality and Nash equilibria. Binmore (2010) view norms as solutions to equilibrium selection problem in coordination games. Gintis (2010) conceptualizes norms as Nash equilibria of supergames that implement correlated equilibria with pure strategies. López-Pérez (2008) sees a norm as a prescription indicating how one ought to behave in any conceivable situation. Complying with the norm triggers emotional consequences which affect players' utility. However, his approach does not address the notion that norms should operate through others' expectations, i.e. the "how others think one ought to behave" aspect that we mentioned above. Bicchieri (2006) discusses several issues, including that norms may transform "mixed-motive" games into coordination games where players however face incentives that depend not only on what others do but also on what they believe.

reduces the weight of normative expectations, relative to the other variables that also influence behavior (personal values and empirical expectations). Hence, depending on the specific content of personal values, individual behavior can become more or less selfish if disagreement about the normative content of the norm increases.

Our study also contributes to the empirical literature on social norms. A large number of experiments, mostly randomized control trials in the field, show how social information can influence behavior in a variety of settings.[5] An interesting recent literature discusses the presence of pluralistic ignorance, defined as incorrect normative or empirical expectations, and shows that correcting it has large effects on behavior (Bursztyn et al. 2020a; Byrne et al. 2018). Individuals are found to look for clues of the prevailing norms from others' behavior and from institutional signals (Tankard & Paluck 2016). Empirical tests of the interaction between normative and empirical expectations show that the two types of beliefs complement each other when consistent (Cialdini et al. 1990), but that empirical beliefs are more influential when at odds with normative ones (Bicchieri & Xiao 2009). The behavioral consequences of aversion to violating empirical and normative expectations are also found to be different (Danilov et al. 2018). We contribute to this literature by focusing on normative expectations, by manipulating different aspects of their distribution, and by examining their impact both on behavior, and on empirical expectations and values.

Our focus on the distribution of normative expectations, and specifically on its variance, also allows us to contribute to the experimental literature on norm consensus and self-serving norm manipulation. Several studies introduce competing social norms with different private returns from compliance, and they typically find that individuals select the norms that best serve one's self-interest (Nikiforakis et al. 2012; McDonald et al. 2013; Bicchieri et al. 2020). We add to these results by showing that reduced consensus around the norm does not necessarily lead to higher prevalence of selfish behavior, and that a form of self-serving bias can also work through increased influence of individual values.

Finally, our study makes a methodological contribution to the elicitation and provision of information on normative expectations. Typically, empirical and normative expectations are elicited and communicated to experimental subjects in the form of observed behavior or in the form of

[5] These setting range from energy and water consumption (Allcott 2011; Ayres 2012; Ferraro et al. 2013; Ferraro et al. 2011), to contributions to charitable causes (Frey & Meier 2004; Shang & Croson 2009), voting (Gerber & Rogers 2009) and financial decisions (Beshears et al. 2015). In a related literature, role models, such as those portrayed by the media or political leaders, have also been shown to have an impact on norm perception and norm-related behavior concerning, for instance, women's status (Jensen & Oster 2009; Beaman et al. 2012), fertility (Chong et al. 2012), or dissent and cooperation (Paluck 2009). Similarly, laboratory experiments show how behavior is influenced by normative information conveyed by leaders' actions (Drouvelis & Nosenzo 2013; d'Adda et al. 2017) or by the institutional environment (Peysakhovich & Rand 2015; Tankard & Paluck 2016).

normative beliefs (e.g., Bicchieri & Xiao 2009). Krupka & Weber (2013) introduce a novel norm elicitation method: they incentivize participants to guess the modal appropriateness rating of each action available to the decision maker. This method leads participants to coordinate over ratings that may not reflect their true normative values. We instead incentivize participants to guess the full actual distribution of individual values (and actions): our approach thus bypasses the coordination problem, as participants' answers should reflect their true guesses about others' values (and actions). Moreover, our approach allows us to empirically examine the role of different moments of the distribution of expectations.

Sections 2-5 describe, respectively, theory, experimental designs, implementation, results. Section 6 concludes.

2. Theory

We next discuss norms and conformity; introduce partiality; restrict attention to one setting for which we specify a formal model; and state precise predictions to be tested.

2.1. Values, conformity and social norms

We start by clarifying what we mean by social norm. In principle, deviations of individual behavior from narrowly defined self-interest can be motivated by three sets of variables: (i) *Values*, namely personal conceptions of what is the right thing to do in a specific circumstance. (ii) *Conformity*, namely a desire to follow the actual behavior of others, such as with fashions. (iii) *Compliance with social norms*, that we define as a desire to comply with the perceived values of others.[6] We reserve the term "social norm" for what is usually called an injunctive or prescriptive norm in the literature, i.e., for something normative, prescribing what one "should" do (Bicchieri 2006; Prentice 2007). Social norms and individual values are related because a certain behavior is perceived as a norm only if there is sufficient consensus that it is the right thing to do.[7]

[6] Departing from a social norm entails an element of disappointing the expectations of others. In this regard, the motivation we look at resembles guilt aversion (see Battigalli & Dufwenberg 2007 for a general model), a belief-dependent sentiment the modeling of which requires the framework of psychological game theory (Geanakoplos et al. 1989; Battigalli & Dufwenberg 2009). However, we consider expectations about how one "ought to behave", not how expectation regarding how one will actually behave, which marks a way that our approach is not formally captured by the papers we cited in this footnote. Recent evidence that attitudes toward conformity help predict which individuals will adhere to or deviate from a social norm supports our view on the importance of conformity and compliance motives (Andreoni et al. 2020).

[7] The element of subjective values is captured by the notion of personal normative beliefs within Bicchieri's (2006) framework, which is however missing from her formal model – see the Appendix of her chapter 1. On the contrary, in our model it is an essential building block.

In practice it may be difficult to disentangle (i), (ii), and (ii), and there may be interaction effects. Nevertheless, for our purposes it is useful to keep (i), (ii), and (ii) conceptually distinct.

2.2. Partial norms

Definitions of norms tend to require consensus. Most accounts require that if *r* (for "right-thing-to-do") is a norm then everyone's *r* is the same, everyone believes everyone's *r* is the same, … et cetera ad infinitum. Call this an *ideal* norm. Beyond philosophical discussion, that concept is implausible. Hardly anything qualifies. Consider tipping in the US, a norm which influences behavior. But is there an ideal norm *r*, in which there is common belief? No. Some people say tipping 15% is a norm, but some say 15-25%, and some add conditions, under which the norm applies, such as "good service." There is no one *r*, and no common belief in any *r*. However, approximate common belief in some (small) set of *r*-values may be good enough to meaningfully talk about a norm. This is what we have in mind when we speak of a *partial* norm.[8]

Once one allows for partial norms it becomes natural to quantify degree of partiality and to explore its behavioral implications. It then seems natural to conjecture that the closer a partial norm is to an ideal one, the stronger is its normative pull on an individual's choice.

2.3. Our setting

We consider the following version of a dictator game. There are two players: player A and player B. Player A is given 10€ and must indicate how much to give to player B, with the understanding that whatever is given will be value-doubled: if A gives x then B gets $2x$ while A gets 10-x. This setting has several virtues. First, it is simple. Second, while in principle norms may prescribe behavior in complicated situations, having multiple active players makes it difficult to infer how norms affect behavior. Suppose a norm (somehow) changes and that *i*'s behavior changes. Is this due to the shift in norm, or is it because *i* believes others' behavior changed? It is hard to tell. A dictator game avoids the problem. Third, the value-doubling feature (player B gets $2x$, not x) introduces some ambiguity over the content of the norm. Without doubling x, a 50/50 split would be an obvious norm. With the doubling, a norm of maximizing the total surplus from the experiment would suggest a choice of x=10. A norm of minimizing inequality would suggest a choice of x=3. If the norm is that A gets to

[8] In principle, a norm can be partial for two reasons: incomplete social consensus, or subjective doubts about the contents of the norm, irrespective of social consensus. We focus on incomplete social consensus while noting that this involves an abstraction.

keep half his endowments, then it would suggest a choice of x=5. We exploit this feature to explore how actions and beliefs react to informational treatments (we return to this point in section 4). [9]
We assume that the player trades off two concerns. On the one hand, his material incentives are such that other things being equal he prefers to keep as much as possible for himself. However, he would also like to behave in a way that is normatively appropriate. If those two goals are incompatible, he has to find a personally optimal choice, which strikes a balance.

What do we mean by "normatively appropriate"? We provide a full answer below, where we also consider norm-partiality as well as a concern for conformity. However, before we go there, let us first consider a benchmark: the case where everyone agrees what the norms says is the right thing to do. That is, all individuals have the same personal values which then coincide with the social norm. Moreover, there is no concern for conformity. In this case we say that there is an *ideal* social norm, which is a number $N \in \{0,1,\ldots,10\}$. Each individual then donates the amount x that minimizes the following quadratic loss function:[10]

$$W = x + (N - x)^2/2\theta \qquad (1)$$

The first term captures the player's material incentives, the second term captures his normative concern: to stay as close as possible to the ideal social norm. The parameter $\theta > 0$ reflects the steepness of this tradeoff. If θ tends to infinity we get selfishness as a limiting case, and the lower is θ the more the player cares about following the norm.

Taking the first order condition of the above minimization problem, one sees that the optimal amount given is the number in $\{0,1,\ldots,10\}$ which is closest to:

$$x = Max\{0, N - \theta\} \qquad (2)$$

Trading off the two concerns that motivate him, the player gives less than N, and how much less depends on the parameter θ.[11]

[9] We conducted a preliminary test with 400 subjects, recruited on the online platform Prolific Academic to test the distribution of r in different variants of the dictator game (DG): a standard DG, a DG where the recipient gets $2x$ as described in section 2.3, a DG such that if player A gives x then player B gets $0.5x$, and a DG with a taking option. The results from this test indicated that the DG of section 2.3 generated the largest variance in individuals' values, r. Full results are available upon request.

[10] The choice of loss function is non-obvious and substantial. See Michaeli & Spiro (2015) for a critical discussion analyzing loss of deviating from a social norm in terms of "curvature of social pressure" in different "societies." They argue that "strict societies are those emphasizing full adherence to the social norm, and hence they utilize concave social pressure; liberal societies are those allowing freedom of expression as long as it is not too extreme, and hence they utilize convex social pressure" (pp. 51-52.) On the presumption that Italy – the place where we run the experiment - is a "liberal society," Michaeli & Spiro's arguments help justify our specification (1).

[11] Note that the typical choice involves some shading relative to N, a typical feature in "liberal societies" following the thoughts of Michaeli & Spiro (2015) (see the previous footnote). This marks a contrast to Michaeli & Spiro's "strict societies,'' and also to models where agents care about their social image and (in equilibrium, for signaling reasons) end up conforming to each other's choices (see Bernheim 1994; Bénabou & Tirole 2006; Andreoni & Bernheim 2009) or to each other's opinions (see Bursztyn et al. 2020b).

Let us now introduce the additional features that are central to our approach. Let $r\in\{0,1,\ldots,10\}$ be a player's view of the "right-thing-to-do" – what above we called *values*. In our analysis r is a primitive notion, underlying much of our analysis. Let $E(r)$ be a subject's expectation of everyone's r – our notion of a *social norm*. Let $V(r)$ be a subject's variance in beliefs of everyone's r – our notion of *partiality* of a social norm. Also actual choices and corresponding beliefs will be important, for modelling players' concerns for conformity. As before, let $x\in\{0,1,\ldots,10\}$ be a player's choice of how much to give. Let $E(x)$ be a subject's expectation of how much others give. Let $V(x)$ be a subject's variance of how much others give. Sometimes we will refer to $E(r)$ and $E(x)$ as a subject's normative and positive expectations, respectively (Bicchieri 2006).

Again, we think of a player as trading off material interests and normative concerns. However, from now on N, which was previously the same given number for all, should be seen, for each individual, as the value of a function, and different values of the arguments may be plugged in for different individuals. More precisely, N reflects the three forces we have previously hinted at: values, social norms and conformity. Specifically, we assume that N is a weighted average of three variables:

$$N = r + \alpha[E(r) - r] + \beta[E(x) - r] \tag{3}$$

where $\alpha, \beta > 0$ and $\alpha + \beta < 1$. This implies that $\alpha, \beta,$ and $1\text{-}\alpha\text{-}\beta$ are, respectively, the relative weights given to normative expectations, $E(r)$, to empirical expectations, $E(x)$ (capturing a conformity concern), and to individual values, r. We have, however, presented N in the (equivalent) form given by (3) in order to emphasize a comparative static that will be put to crucial use below: If $E(r) > (<)\ r$, then N is increasing (decreasing) in α, the weight on the social norm $E(r)$. Importantly, as we discuss below, the weights α and β are fixed for any given individual, but their magnitude may reflect the properties of the distribution of values r in the population (more on this below).

For simplicity of exposition, we assume throughout that players differ only in their values, i.e., they have different realizations of the random variable r (and hence N) but they have the same preferences and the same parameters α, β, θ. Note that if α tends to 1 and if $V(r) = 0$, then for all individuals it holds that $r = E(r)$ and their N's would be the same. In the remainder of the analysis, we assume that individuals have different values of r and that $1 < \alpha < 0$.

Under rational expectations and common knowledge, equation (3) implies:

$$E(N) = E(r)(1 - \beta) + \beta(E(x)) \tag{4}$$

In words, $E(N)$ can be interpreted as capturing the influence of expectations on the normative drivers of behavior. By (4), $E(N)$ is a weighted average of normative ($E(r)$) and empirical expectations ($E(x)$), where β is the weight given to empirical expectations.[12] We solve for the full rational expectations equilibrium under the assumption that the non-negativity constraint on x is never binding, for all realizations of r and for any value of $E(x) \geq 0$. This assumption is satisfied if the weight β on the expectations of others is not too large, specifically if

$$\beta < Min\{1 - \frac{\theta}{E(r)}, \frac{(1-\alpha)\underline{r} + \alpha E(r) - \theta}{\underline{r}}\}$$

where $\underline{r} < E(r)$ denotes the smallest possible realization of the random variable r. Under this condition and under common beliefs, the equilibrium is unique and by (1),

$$E(x) = E(N) - \theta > 0 \tag{5}$$

The full rational expectations equilibrium is obtained by solving the system of linear equations (2)-(5). The equilibrium amount given, as a function of own values and of the norm is:[13]

$$x = r + (\alpha + \beta)[E(r) - r] - \frac{\theta}{1-\beta} \tag{6}$$

We thus have that in equilibrium the amount given by each player is an increasing function of his expectations of the normative beliefs of other players, $E(r)$, and of his own values, r. Moreover, if individuals place more weight on the expected normative standards of others (i.e., if α increases), then the amount given rises (falls) if $E(r) - r > 0$ (< 0). Intuitively, if $E(r) - r > 0$ then the individual expects others to be more generous than he is. Putting more weight on the normative standards of others induces him to give a greater amount, but the opposite happens if he perceives others to be less generous than he is. Similarly, a higher weight β on the expected actual behavior of others has an ambiguous effect, that depends on the sign of $E(r) - r$ but also on the size of θ (the weight given to selfish utility as opposed to normative concerns).

We re-emphasize that the only source of variation across individuals is in their values, r. Specifically, the variance in the amounts donated across individuals is $V(x) = (1 - \alpha - \beta)^2\ V(r)$ where $V(r)$ denotes the variance of r. Thus, putting more weight on normative and/or positive expectations (i.e., increasing α and/or β) also reduces the variance $V(x)$ of the amounts given by different players. Intuitively, higher weights on expectations reduce the weight on the only idiosyncratic variable that varies across individuals.

[12] The evidence in Bicchieri & Xiao (2009) can be interpreted as indicating that β is close to 1 – only empirical expectations influence behavior. But we think that such a strong interpretation of their results is not warranted. Even in the context of their experiment, information about the actual behavior of others can be interpreted as a credible signal of the value of others (i.e. as shifting both $E(r)$ and $E(x)$).

[13] If the condition on β is not satisfied, then a rational expectations equilibrium still exists and also admits $x = 0$, although it need not be unique and to obtain a closed form solution we need to impose a specific functional form on the distribution of r.

Finally, we address the key issue of partiality of norms. Although we have written the relative weights α and β as parameters, they are likely to reflect other features of the environment. In particular, it is reasonable to conjecture that the size of α is also affected by how much consensus there is around the value system captured by the random variable *r.* As emphasized also by Bicchieri (2017), the more consensus there is, the more relevant is the desire to conform to what others regard as good behavior. In our context, this can be captured by the following assumption:

The relative weight α is a decreasing function of V(r) *(A)*

In the limit, if everyone shares the same value system, then there is an ideal norm equal to *E*(*r*). As hinted at before, in this case *E*(*r*)=*r* since the player himself is included in "everyone." Individual values and social norms coincide. If instead individuals have very different value systems, then the subject would no longer believe that an ideal norm existed. Rather, the norm would be partial, still equal to *E*(*r*), but less likely to influence behavior. Thus, the relative weight α would be smaller.

Under assumption (A), therefore, a higher variance *V(r)* increases the dispersion in the amount given, *V(x)*, both directly and indirectly (through a lower value of α). The effect on the amount given, however, is heterogeneous: recalling the previous discussion, a higher variance (and hence a lower value of α) increases (decreases) *N*, and thus it increases (decreases) the amount given if $E(r) - r > 0\ (< 0)$. In other words, when partiality increases an individual derives his subjectively perceived *N* by putting more weight on his own values, *r*, than on the values of others, *E(r)*. If his values are more altruistic than the average (i.e., if *r>E(r)*), his *N* will increase thus he will be more generous. If *r<E(r)*, his *N* will decrease thus he will behave less generously.

The result that a more partial (and hence less influential) social norm does not always lead to more selfish behavior may seem puzzling. One might expect that normative disagreement induces individuals to put more weight on their selfish motives through self-serving bias, and hence to reduce the amount given (see for instance Nikiforakis et al. 2012). The intuition for our result is that altruistic behavior is not only driven by the perceived social norm, but also by subjective values (the variable *r*); normative disagreement reduces the relative weight on normative expectations, and increases the relative weight on subjective values. If the norm is more partial, and hence less influential, behavior is driven to a larger extent by individual values, that can be more or less demanding than the social norm. Hence, a weaker social norm certainly leads to less conformity in behavior, but its effect on donations is heterogeneous and depends on whether the individual's values are more or less altruistic than the average.[14]

[14] Note, however, that we have implicitly assumed that the tradeoff between selfish and normative concerns is not affected by consensus; i.e., parameter θ in equation (1) does not depend on *V(r)*. If it did and the relationship was negative, there would be an additional implication that would tend to make individuals more selfish as *V*(*r*) rises.

2.4. Summing up

The analysis of section 2.3 suggests that individual behavior can be influenced by four distinct dimensions:

- a value-per-se dimension: individual beliefs of what the right thing to do is, i.e. r;
- a norm-compliance dimension: individual beliefs about others' r's, i.e. $E(r)$;
- a conformity dimension: individual beliefs about x, i.e. $E(x)$;
- the degree of consensus around the value system, reflected in the weight α.

We make the following key predictions:

- the amount donated, x, is increasing in the first three dimensions, r, $E(r)$ and $E(x)$;
- less consensus on the right thing to do, measured by a higher variance *V(r)*:
  (i) increases the variability across individuals of the amounts donated, *V(x)*;
  (ii) decreases (increases) the amount given if $E(r) - r > 0\ (< 0)$.

Finally, norms may influence not only behavior but also beliefs. A direct implication of our model is that, under rational expectations, a higher $E(r)$ induces and increase in $E(x)$ - cf. (4) and (5). Moreover, through social learning, higher *E(r)* could also affect individual values and raise *r*.

In section 3 we describe the experiment that we designed to test these predictions. Many of key features of the theory have design counterparts that create relevant *exogenous variation*. Where the theory calls for $E(r)$ or extent of disagreement, *V(r),* to influence x, our design includes some feature that induces exogenous (by treatment) variation in $E(r)$ or in *V(r).* However, some of the relations predicted by this model involve correlations between two endogenous variables (e.g., between amount given, *x,* and positive expectations, *E(x)*). Our ability to test the relations between these jointly endogenous variables is limited, as further discussed below.

3. Experimental designs

We randomly form pairs of subjects, one of which is randomly assigned the role of "Individual A" while the other one plays in the role of "Individual B". The former is the *dictator* (a term that we never use in the instructions). He/she is given €10 and has to choose how much to transfer to individual B, knowing that the transfer will be doubled.

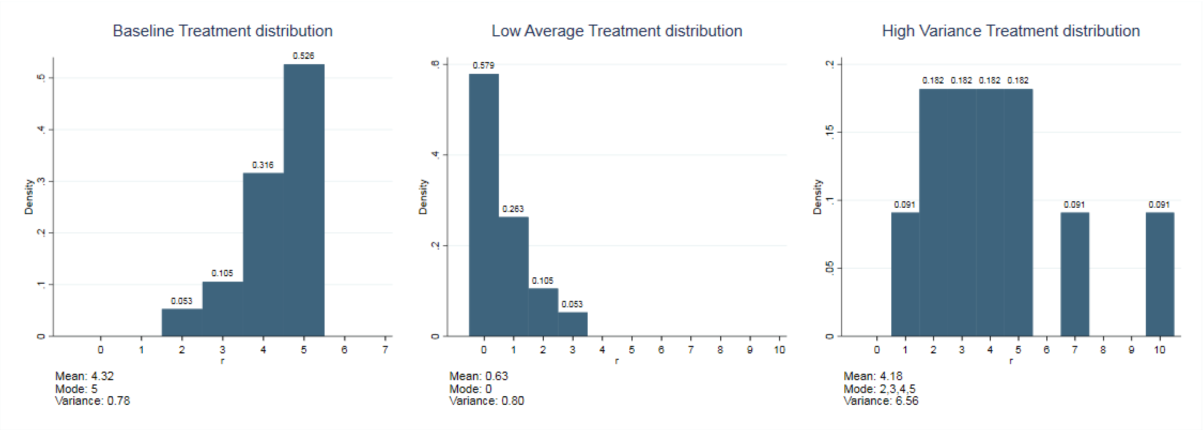
Our design achieves exogenous variation in individuals' beliefs through three *Information Disclosure Treatments* that vary what other people think of the right-think-to-do. Different information is disclosed in the form of the following distributions:

1. **Baseline**. The baseline distribution is in the left-hand diagram of Figure 1: its mean is equal to 4.32, its mode is equal to 5 and its variance is equal to 0.78.

2. **Low Average**. Subjects are shown the distribution in the center diagram of Figure 1. Relative to the Baseline distribution, its variance is not statistically different, but its mean is statistically lower, equal to 0.63.

3. **High Variance**. We show subjects the distribution in the right-hand diagram in Figure 1. Relative to the Baseline distribution, its mean is statistically the same, but the variance is significantly larger.[15]

Figure 1. Treatment distributions



Note: We elicit subjective values, *r*, by asking participants in a pilot study the question "Which is the most appropriate action that Individual A should take?" Each treatment displays the distributions of different sub-samples of answers to this question. For more details on the belief elicitation and pilot study, see Sections 4.2 and 4.3.

Table 1 summarizes the average, the mode, and the variance of each distribution.

Table 1. Treatment distributions

Treatment	Mean	Mode	Variance
Baseline	4.32[a]	5	0.78[b]
Low Average	0.63	0	0.80[b]
High Variance	4.18[a]	2,3,4,5	6.56

[a] The means of Baseline and High Variance distributions are not significantly different from each other (p-value: 0.87).
[b] The variances of Baseline and Low Average distributions are not significantly different from each other (p-value: 0.96).

[15] We also ran one other informational treatment with a milder contrast to the Baseline. Specifically, we presented subjects with a distribution such that the Low Average Treatment had an average value for *E(r)* = 3.23, which was significantly lower than that in the Baseline treatment. The variance was *V(r)* = 0.71, not statistically different from that in the Baseline treatment.

As we disclose information, we tell participants that we collected other subjects' opinions in previous sessions of a similar experiment. We leave participants uninformed of the non-random selection the distribution, with the intention of changing their beliefs. We are careful in not suggesting participants that the reported distributions represented a general pattern.[16]

We run two versions of our experiment, one with Actions (hereafter, Experiment A), and another with Belief elicitation & Actions (Experiment B&A).[17] The rest of this section explains.

3.1 Experiment A

In Experiment A, dictators choose their donations immediately after information disclosure. This design cleanly and directly tests how information about the social norm affects actions. A higher amount donated in the Baseline treatment than in the Low Average treatment would imply that subjects' generosity is *causally* affected by the generosity of the social norm. Thus, the comparison between the amounts donated in the two treatments represents a clean test of the prediction that the amount donated, x, is increasing in $E(r)$.

Moreover, a higher dispersion of the amounts given in the High Variance treatment compared to Baseline would support the theoretical prediction that norm partiality also influences behavior. However, Experiment A cannot test the more precise prediction that the effect of $V(r)$ on the amount given by each individual depends on the contrast between own values r and the social norm $E(r)$, as we do not observe these individual beliefs. We can do it instead using beliefs elicited in Experiment B&A, although this raises other issues discussed below.

The identification of treatment effects on actions comes at a cost, as this design only tests a reduced form of our theory of norm compliance and conformity. Experiment A enables us to observe whether $E(r)$ or $V(r)$ affect behavior, but we cannot pin down the channel. For instance, it is plausible that x is affected by $E(r)$ directly, or it is affected indirectly, because a change in $E(r)$ may lead to a change in $E(x)$ or in r. As for $V(r)$, there might be multiple channels as well. A higher $V(r)$ may plausibly lower the desire to comply with $E(r)$ – a lower value of parameter α, as postulated in our theoretical model. It may also lower the desire to conform with $E(x)$ – as captured by a lower β. Moreover, we cannot exclude that a change in $V(r)$ may also lead to a change in $E(r)$ or $E(x)$ and then affect behavior through a change in subjects' beliefs about others.

[16] Formally, there is no deception involved in our experiment, since none of the information provided is false or incorrectly described. Bicchieri & Xiao (2009) adopt similar non-deceptive manipulation in informational treatments.

[17] Experimental instructions are reported in Appendix B.

3.2 Experiment B&A

In order to shed more light on the channel through which treatments affect behavior, in Experiment B&A we prompt the beliefs of all participants: we elicit three different beliefs: a) a subject's belief about the right-thing-to-do, *r*; b) his/her beliefs about others' beliefs; c) his/her beliefs about the amount donated by others (see next section for the incentivizing scheme). We do it before participants are revealed their role in the game. Then, we assign roles and ask dictators to make their donations. Our treatments provide subjects with different information about others' opinions of the right-thing-to-do, *r*. Thus, experiment B&A allows us to study whether subjects' beliefs are *causally* affected by information about the social norm, *E(r)*.

A decrease in subjects' perception of what is right-thing-to-do, *r*, in the Low Average treatment would support the hypothesis that people think they should be less generous if more people think the same. In other words, *r* is positively affected by *E*(*r*). We can also test the hypothesis that positive beliefs, *E*(*x*), are positively affected by *E*(*r*). In this case the Low Average treatment should induce a drop in *E(x)*.

The treatment High Variance is intended to exogenously increase a subject's beliefs about the *partiality* of a norm. We can test whether this treatment induces higher dispersion in normative or positive beliefs, *V*(*r*) and *V*(*x*)*,* respectively.

Experiment B&A could reveal the mechanism through which norms affect individuals' donations, by showing how both actions and beliefs respond to the information treatments. In practice, however, there are several reasons to doubt that we can detect changes in dictators' actions through this design. First, belief elicitation induces players to focus their mind on the subject matters asked in the questions, and in particular to reason on what the-right-thing-to-do is (before knowing whether they will act as dictator or recipient). Second, the design introduces a long time span between the informational treatment and the choice of actions. Third, belief elicitation requires a considerable amount of cognitive effort and causes cognitive fatigue, which might independently affect later choices. These features of the design may dampen the effect of the information on dictators' actions. In particular, reasoning in the abstract on "what is the right thing to do" and on how others respond to this question may lead participants to give more weight to the intrinsic merit of the alternatives faced by the dictator. This could lead them to act based on moral or value criteria (i.e. consistently with what they said is "the right thing to do"), discounting social conventions and hence reducing the effect of information about how others perceive the social norm. For these reasons, in what follows we use Experiment A to test our predictions on the treatment effects on actual behavior, and we use Experiment B&A only to assess the treatment effects on beliefs. [18]

4. Implementation

The sessions were conducted at BELSS *(Bocconi Experimental Laboratory in Social Sciences)* in Milan, during September 2017-February 2018. We recruited 686 students.[19] No subject participated more than once. We ran 15 sessions of Experiment B&A and 16 sessions of Experiment A.[20] The average number of participants per session was 19. Sessions lasted on average an hour for Experiment B&A and 30 minutes for Experiment A. The average payment was €9.54 including show-up fee.

Table 2 summarizes the overall setup.

Table 2. Experimental conditions and number of subjects per condition

	Experiment B&A		Experiment A	
	Number of subjects	Number of sessions	Number of subjects	Number of sessions
Baseline	98	5	98	5
Low Average	96	5	92	5
High Variance	96	5	114	6

Note: Within each cell, data on DG allocations are available only for the subjects assigned to the role of dictators, while we have data on beliefs for all subjects in the B&A treatments.

The experiment and payment protocols were designed to ensure the highest degree of anonymity and minimize the possibility that subjects' choices were driven by reputational concerns or experimenter demand effects. Upon arrival, *n* subjects entered the experimental laboratory one by one and were randomly assigned an isolated seat with a computer terminal. The number of participants in a session, *n*, was always an even number between 16 and 20.[21] Participants could read instructions on the computer screen. At the beginning of each session, an assistant read aloud the General Instructions (see below) and checked that participants correctly understood them. The experiment was conducted with real money.

[18] Bicchieri (2006) and Cialdini et al. (1990) also argue that situational cues that make a certain social norm salient increase norm-compliant behavior. Such situational cues include discussing or thinking about a social norm. The focusing effect of social norms, i.e., that simply thinking about a social norm causes norm compliance, is empirically confirmed by Krupka & Weber (2009).

[19] Subjects were recruited from the Laboratory's sign-up list using ORSEE (Greiner, 2015).

[20] We also ran 5 sessions of a Pilot Experiment, that we describe in section 4.3 below. We did it in order to generate data to run the Baseline informational treatment.

[21] In order to reduce the risk of having too few people in a session, we recruited 22 people. If for instance 22 participants showed up, we randomly selected the 21th and the 22nd participant to exclude from the session. If say 17 people (an odd number) showed up, we randomly chose the 17th to exclude. All excluded persons were paid an increased show-up fee of €5 and were allowed to sign up for another session in the future.

4.1 Experiment A

4.1.1. Phase 1: General instructions

Subjects were informed about the number of participants *n*. We told them that another participant in the room had been randomly paired with them. Thus *n*/2 pairs had been formed. One subject in the pair would soon be randomly assigned the role of "Individual A" while the other one would be assigned the role of "Individual B". All subjects found a carton box on their desk. We asked them not to open it until instructed to do so. We informed them that the box of Individual A contained two small bubble cushioned envelopes, one white and one yellow. The former was marked "Money for Individual A", and it contained €10 in one-euro coins. The latter was marked "Money for Individual B", and it was empty. The subject playing in the role of Individual A would be able to choose how many coins x to transfer to Individual B, by simply putting them in the yellow envelope ($x \in \{0,1,\ldots,10\}$). Individual A would keep the remaining coins while the experimenters would take care of the transfer to Individual B. They would also match the transfer with additional money. Thus there was common understanding that, by the end of the experiment, individual B would receive $2x$. All participants were informed that, whoever Individual A was, no other person would ever know his/her identity and choice. Hereafter participants could read instructions on the computer screen.

4.1.2. Phase 2: Information disclosure

Before participants were assigned their role, the computer showed them a distribution of other participants' opinions about the most appropriate donation amount.[22] Participants were informed that those opinions had been previously gathered in sessions similar to the one they were in. We had three treatments as described in section 4: Baseline - Low Average - High Variance (cf. Figure 1). This is when our information manipulation eventually kicked in.

4.1.3. Phase 3: Role assignment, actions and transfers

Participants could now read on screen the role they were randomly assigned in the pair. They could open the box. Those who were assigned the role of Individual B found it empty. Then they were instructed by the computer to remain silent. Those who played in the role of Individual A found the two padded envelopes. The computer instructed them to silently transfer their donation x from the white envelope to the yellow envelope for Individual B, and put the remaining amount in their pocket. Their actions could not be seen by anyone. No one could infer from the noise they eventually made whether they were putting coins in their pockets or in the envelope of Individual B. After making

[22] Participants were shown the distribution of others' answers to the question: *In your opinion, which is the most socially appropriate action that Individual A should take* (cf. Figure 1). We made it clear that by "socially most appropriate" we meant behavior that they considered the "correct" or "ethical" thing to do.

their actions, all Individuals A were asked to leave the two envelopes in the box and record their choice on the computer. Then experimenters collected all of the boxes (also those of Individuals B) and took them to another room, while all subjects remained seated and silent. In the other room, yellow envelopes containing transfers were actually transferred from each A's box to his/her paired B's box. All of the boxes were returned to the participants. Receivers could then keep the amount received and check if it was the correct amount shown on the computer screen.[23] They were told that that amount would be doubled at the end of the experiment. This procedure ensured complete anonymity between subjects. No subject in the room could infer the identity of his/her paired subject or whether any other participant was an Individual A or an Individual B. *4.1.4. Phase 4: Payments* Participants could read on the screen the total amount they earned, including the participation fee. Then subjects were called one by one outside of the room by their seat number. They received an anonymous envelope with their seat number marked on it. The envelope contained the money they earned in addition to the coins they already got during the experiment. This procedure was designed to ensure the maximum degree of anonymity during the experiment, and to minimize the risk that dictators made their choices to please the experimenter.

4.2 Experiment B&A

Relative to Experiment A, Experiment B&A includes an additional Phase 2a, in which we elicit beliefs. All other phases are the same.

4.2.1. Phase 2a: Belief elicitation

After disclosing information just like in Experiment A, we asked three questions. First, "Which is the most appropriate action that Individual A should take?" In other words, we asked subjects to tell us their "*r*'s". They could tick one of the eleven boxes in a table containing the eleven possible transfers. Second, we asked them to guess the distribution of the answers that the *n* participants in the same session gave to the first question. They had to guess eleven frequencies, which we incentivized by paying €0.2 for each right guess. From the answers, we compute each participants *E(r)* and *V(r)*. Third, we asked them to guess the distribution of the actions that the *n/2* Individual A's would take in the same session. Also in this case they had to guess eleven frequencies, which we incentivized with €0.2 for each correct guess. From the answers, we compute each participants *E(x)* and *V(x)*.[24]

[23] We also asked each receiver for her feelings about the dictator's allocation. Receivers could say by how much they would be willing to reduce Individual As' earnings: this decision was completely hypothetical and unincentivized. Receivers knew that their payment reduction decision would not be implemented, thus their expressions of approval or disapproval had no real consequence.

[24] In phase 4 of Experiment B&A (Payments), besides other payments, participants could read on the screen also rewards from guessing other participants' answers or behavior.

4.3 Pilot

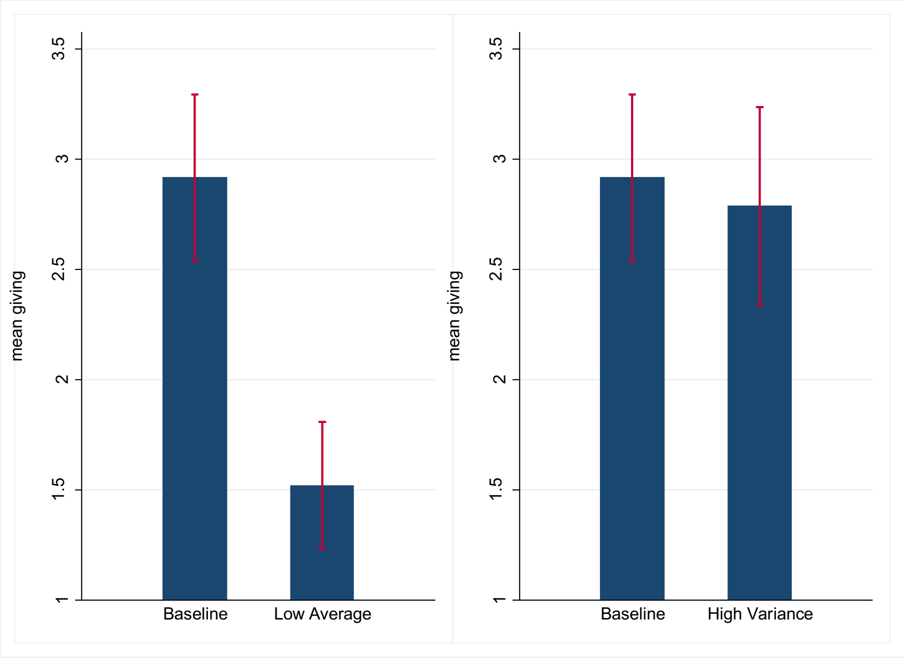
We also ran a Pilot experiment to test the design of the Experiment B&A. We used the data on individuals' opinions about *r*, the right-thing-to-do to selectively build the distribution that we used in the Baseline treatment of both Experiment B&A and Experiment A. The incentivizing scheme and the payment procedures are the same as in Experiment B&A.

5. Results

5.1. Treatment effect on actions

We begin by focusing on experiment A, and explore how actions are influenced by the informational treatments. We comment on experiment B&A in the next section. Table A1 in the Online Appendix A provides summary statistics of average dictator giving in Experiment A and B&A, by treatment (the overall amount given on average is € 2.29).

Figure 2. Dictator giving in Experiment A, mean-comparison test

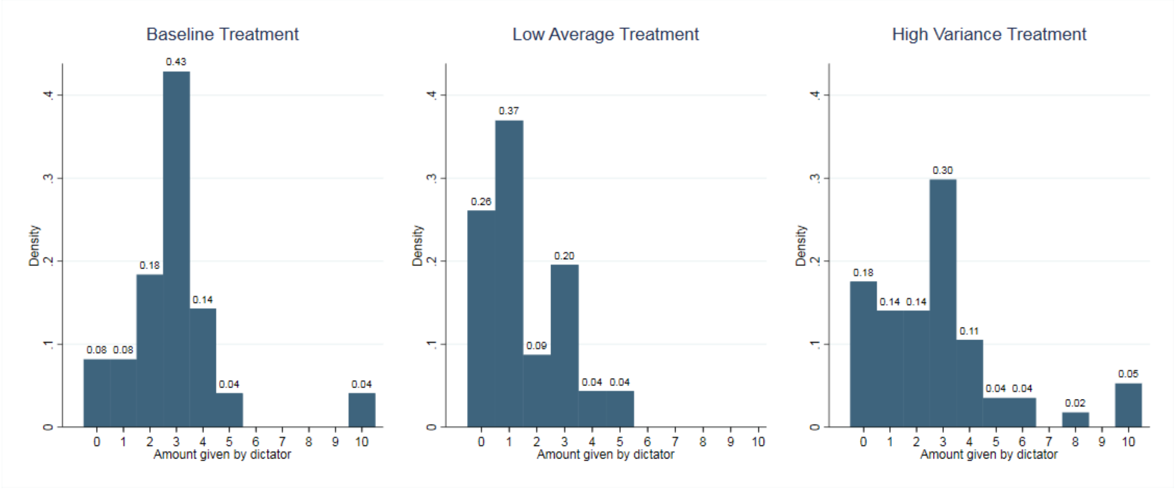


Notes: Bars denote means, whiskers 95% C.I. Wilcoxon rank-sum test results: left Prob > |z| = 0.0000; right Prob > |z| = 0.2120.

Figure 2 shows the average amounts given by the dictator, in a pairwise comparison, by treatment. As shown in the left hand panel of Figure 2, the difference in the average amount donated between the Baseline and the Low Average is about €1.4 and is statistically significant (Wilcoxon rank-sum test: p = 0.0000). This corresponds to almost 50% of the average donation in the Baseline treatment, and to about 40% of the difference in *E(r)* across the two treatments (in the Baseline treatment participants are shown a distribution that has *E(r)* = 4.36, while in the Low Average treatment they are shown a distribution with *E(r)* = 0.63). As shown in the right hand panel of Figure 2, there is no significant difference in the amount donated between the Baseline and the High Variance treatments (Wilcoxon rank-sum test: p = 0.2120). Both results are in line with the predictions of the model.

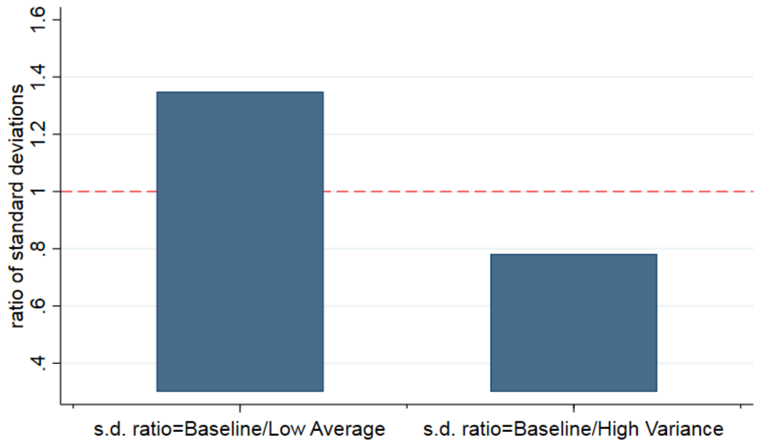
In order to assess whether the treatments also affected the variability of donations, we turn to the distribution of dictators' choices. Figure 3 illustrates the histograms of the actions chosen by the dictator under the three treatments. The three distributions look different, particularly the first two. Compared to the Baseline, the distribution of actions in the Low Average treatment is shifted towards 0, while the distribution in the High Variance Treatment seems more spread out, in line with the predictions of the model. Nevertheless, statistical tests only confirm one of these two visual impressions. Specifically, Pearson's Chi-square distribution tests reveal that the distribution of giving in the Baseline treatment is significantly different from that in the Low Average treatments (p = 0.001), but not from the one of the High Variance treatment (p = 0.339).

Figure 3. Distribution of dictator giving in Experiment A, by treatment



The theory predicts that the dispersion of the amounts given is increased by the high variance treatment, but not by the low average treatment. Figure 4 compares the standard deviation of the distribution of the amounts given by the dictators in each treatment. Given that donations do not appear to be drawn from a Normal distribution, we use Levine's robust test statistics for the equality of variance to assess the statistical significance of the differences in the variance of giving across treatments. The High Variance treatment indeed yields a larger variance in the amounts given than the Baseline treatment, with a p-value of 0.07, while the Low Average treatment displays lower standard deviation than the Baseline treatment, but this difference is not statistically significant.

Figure 4. Dictator giving in Experiment A, ratio of standard deviations



The findings illustrated in Figure 2 are confirmed also by simple linear regressions. We regress the amount given on dummy variables for the Low Average and High Variance treatments. Only the Low Average treatment is statistically significant, as expected, and the estimated coefficients are quite stable and similar to the effects displayed in Figure 2.[25] Distinguishing between the probability to give a positive amount, and the amount donated, conditional on giving, regression analysis shows that the former is negatively affected both by the Low Average and the High Variance treatments, although with a lower level of statistical significance in the latter case. Conditional on giving, the amount given is significantly smaller only in the Low Average treatment.[26] These results are reported in Online Appendix Tables A2 and A3.

Overall, these findings suggest that being exposed to information about the normative beliefs of others affects individual behavior. On average the dictator is more generous, both on the extensive and on the intensive margin, if he is told that more people consider that a more generous behavior is socially appropriate. More dispersion in the normative beliefs of others increases the dispersion of individual donations, but has no effect on the amount given (the fraction of individuals

[25] We obtain similar results when we compare average donations in the Baseline versus the milder Low Average Treatment (see footnote 15): average donations in the milder Low Average Treatment are smaller than in the Baseline, and the difference is statistically significant (Wilcoxon rank-sum: p = 0.0164).

[26] Note that, to obtain a simple closed form solution, the theory has been solved for a linear model, imposing restrictions on parameter values such that in equilibrium everyone always gives a positive amount irrespective of his subjective values. Relaxing these restrictions would imply that some players could give 0. The solution described by equation (6) would no longer hold, but a lower normative expectation *E(r)* would increase the likelihood of observing 0 as well as the average amount given conditional on giving, in line with the empirical findings of the Low Average treatment.

who give 0 goes up in the High Variance Treatment, but this could simply reflect the increased dispersion of the amounts given).[27]

These results are consistent with our theoretical priors. Being exposed to information about what others regard as socially appropriate changes behavior in the direction of the average perceived social norm. Moreover, being informed that there is more disagreement over the contents of the social norm leads to more dispersion in individual donation but has no effect on the average amount given. With these data we cannot test the more specific predictions on the heterogeneous effects of the high variance treatment.

We now explore the mechanisms behind these effects.

5.2. Treatment effect on beliefs

We use data from Experiment B&A to study how informational treatments affect individual beliefs, focusing in particular on: (i) individual assessments of what is the-right-thing-to-do, *r;* (ii) individual beliefs of what others regard as the-right-thing-to-do on average, *E(r)*; (iii) individual beliefs of what others will actually do on average, $E(x)$. Since for (ii) and (iii) we observe the whole distribution of beliefs, we can also study the treatment effect on the variance of normative and positive beliefs of each respondent, namely $V(r)$ and $V(x)$. Recall that this sample of respondents is different from that analyzed above in section 5.1 (i.e., Experiment A), where play of the dictator game was preceded by the informational treatment but there was no belief elicitation.

Table 3 presents descriptive statistics of the three beliefs by treatment. Consider first the comparison over the contents of beliefs, i.e., compare the elements of each row. For all treatments, $E(x)$ is always smaller than both r and $E(r)$ (all p-values = 0.0000, Wilcoxon signed-rank tests). In other words, on average individuals expect that dictators will donate less than what they deem the right thing to do, and of the perceived social norm. This is consistent with our model and with the ideas in the literature (e.g., Bicchieri 2006).[28]

Next, consider the treatment effects (i.e., compare the element of each column). Compared to the Baseline, the Low Average treatment displays significantly lower beliefs in all three dimensions, r, $E(r)$ and $E(x)$ and the differences are statistically significant (all p-values ≤ 0.0001, Wilcoxon rank-sum tests). The effect of the Low Average Treatment is particularly strong on $E(r)$ and $E(x)$. In the High Variance treatment, instead, beliefs are not statistically different from the Baseline. Note also

[27] We find no significant treatment effects on our hypothetical measure of receivers' aggrievement (see footnote 23). Results are available upon request.

[28] In our model the expected value of (6) yields $E(x) < E(r)$, where $E(r)$ denotes both the perceived social norm (second order beliefs), but also average values (i.e. what the average individual deems the right thing to do).

that the standard deviation of *r* is highest in the High Variance treatment, and higher in the Low Average treatment than in the Baseline.

Table 3. Average beliefs in Experiment B&A, by treatment

	N	*r*	*E(r)*	*E(x)*
Baseline	98	3.643	4.020	2.347
		(1.151)	(0.916)	(1.525)
Low Average	96	3.094	1.897	1.264
		(1.693)	(1.287)	(0.967)
High Variance	96	3.698	4.003	2.725
		(1.795)	(1.214)	(1.635)

Note: the table reports means and standard deviations, in parentheses, of subjects' elicited beliefs in Experiment B&A.

The fact that different components of beliefs react to treatments in the same direction reflects a strong positive correlation between such components. Table 4 shows that a subject generally believes that, on average, others share his/her own assessment of the right-thing-to-do, as shown by the Spearman correlation coefficient between *r* and *E*(*r*) (ρ = 0.491, p = 0.000). He/she also expects that others' actions will be consistent with what she regards as the right-thing-to-do, as captured by the correlation between *r* and *E*(*x*) (ρ = 0.363*,* p = 0.000). Finally, a subject expects that others' actions will be consistent with what they regard as the right-thing-to do: the correlation coefficient between *E*(*r*) and *E*(*x*) is ρ = 0.540 (p = 0.000).

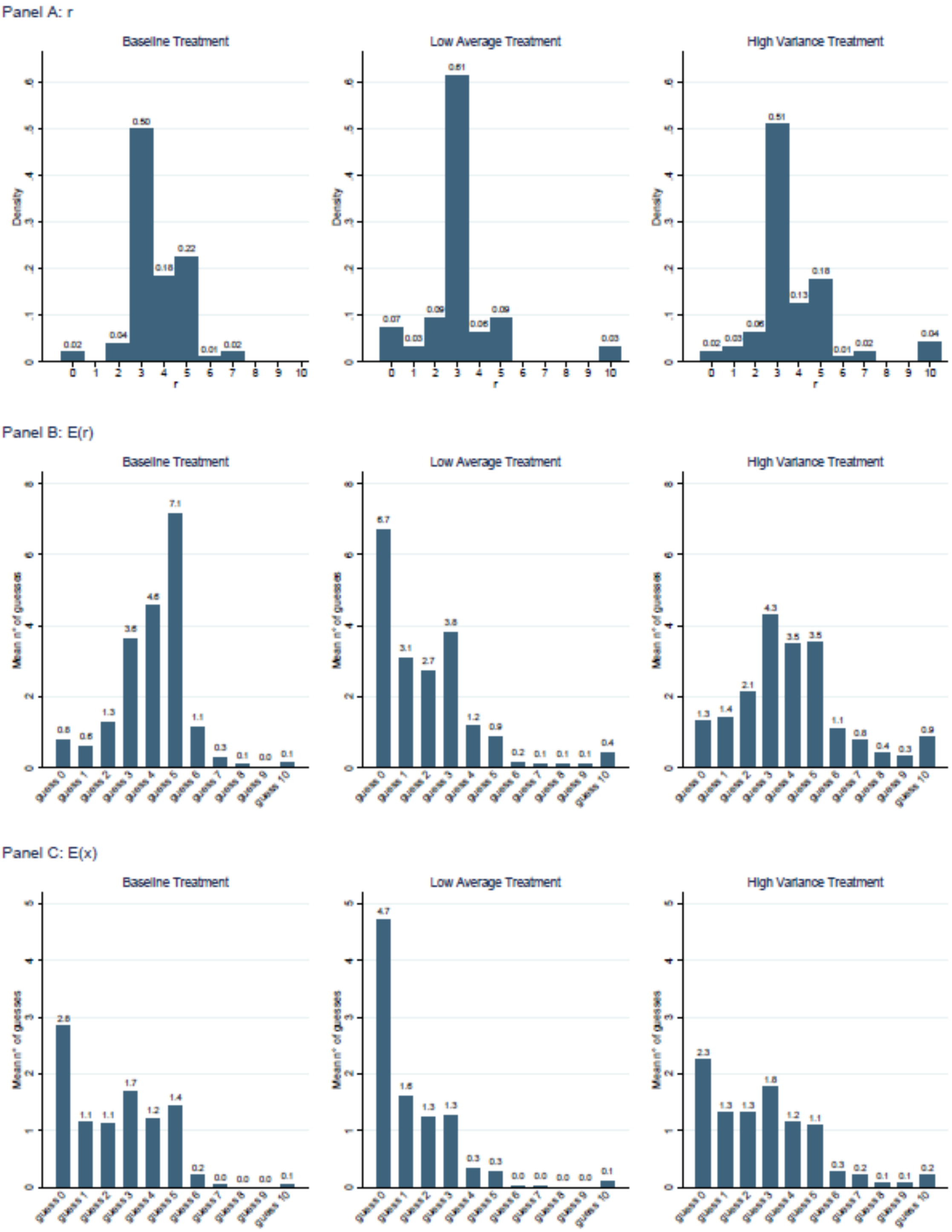
Table 4. Correlation between *r*, *E(r)* and *E(x)*

	r	*E(r)*	*E(x)*
r	1		
E(r)	0.491***	1	
E(x)	0.363***	0.540***	1

Note: Spearman correlation coefficients. Baseline, Low Average, and High Variance treatments are included. *** p<0.01, ** p<0.05, * p<0.1.

We now exploit our rich data on beliefs to examine their distributions. Figure 5 depicts the distribution of *r* in Panel A, of *E*(*r*) in Panel B and of *E*(*x*) in Panel C. The distributions look quite different across treatments, in particular the Low Average treatment is associated with a shift of the distribution of beliefs to the left, towards lower values, as expected. Pearson's Chi-square distribution tests reveal that the distributions of all types of beliefs in the Baseline and the Low Average treatments are significantly different (p = 0.001 for r, p = 0.000 for $E(r)$ and p = 0.004 for $E(x)$). There instead are no significant differences in the distributions of beliefs between the Baseline and the High Variance treatments, consistently with what we found in the previous section for the distributions of the amounts given.

Figure 5. Distribution of beliefs, Experiment B&A



Regression analysis confirms these results (Online Appendix A Table A4). Being exposed to the Low Average treatment induces a drop in one's perception of what is the right-thing-to-do, *r,* in beliefs of what others regard as right-thing-to-do on average, $E(r)$, and in the positive beliefs of what others will actually do on average, $E(x)$. The High Variance treatment increases the dispersion of normative and positive beliefs, $V(r)$ and $V(x)$, with no effect on average beliefs (except for a small increase in $E(x)$ which is only significant at the 10% level).[29]

Since all beliefs move in the same direction with the treatment, we cannot fully disentangle which beliefs are responsible for the observed change in behavior. Nevertheless, the effect of the Low Average treatment on $E(r)$ is particularly large, relative to the effects on $E(x)$ and r respectively. This suggests that the Low Average treatment affects behavior mainly through its impact on individual beliefs of what others regard as socially appropriate (i.e., on the content of the social norm, *E(r)*), rather than through an effect on individual values, r, or on expectations of actual behavior of others. This result is consistent with the mechanism postulated by our theory.[30]

Finally, in the sessions of experiment B&A, we also asked participants to play the dictator game after expressing their belief. We can thus explore how the informational treatments influenced actions in this setting too, comparing each treatment to the Baseline. Unlike in Experiment A (see Figure 2 above), here neither treatment has any effect on the average amount given, compared to the Baseline (Figure A1 in the Online Appendix A). The High Variance treatment increases the dispersion in the amount given compared to the Baseline, as in Figure 3, but the variance of amounts given is also larger in the Low Average treatment than in the Baseline (Figure A2). These results are confirmed by OLS estimates (Tables A7 and A8). Thus, as argued above in section 3.2, the belief elicitation stage dampens the treatment effects on dictator giving. When we add belief elicitation to the experiment, the evidence of a causal effect of information on the amount given disappears.

This also explains why we fail to find support for prediction (ii) on the heterogeneous effect of a change in *V(r)*. Recall that we predict that more consensus on the right thing to do increases (decreases) the amount donated if $E(r) - r > 0$ (< 0). Testing this prediction requires examining whether the impact of the high variance treatment on donation differs between subjects with $E(r) > r$ and subjects with $E(r) < r$. We thus can test this prediction only with data on both beliefs and

[29] We obtain similar results when comparing beliefs in the Baseline and in the milder Low Average treatment (see footnote 15). The milder Low Average treatment has the expected effect on beliefs on what others perceive to be the right-thing-to-do: *E(r)* is significantly lower (p = 0.000, Wilcoxon rank-sum test) than in the Baseline treatment. In the milder Low Average treatment beliefs on others' actions, *E(x)*, are also on average lower, but not significantly so (p = 0.4809), while *r* is higher on average, also not significantly (p = 0.8814).

[30] Tables A5 and A6 in the Online Appendix repeat the same exercise controlling for time of day or day fixed effects. Some of the estimated coefficients lose significance, but the effect of the Low Average treatment on *E(r)* and the effect of the High Variance treatment on *V(r)* and *V(x)* are robust and stable.

actions for the same subjects, i.e. with data from Experiment B&A. However, the lack of effect of the experimental treatments on actions in Experiment B&A limits our ability to perform this analysis. Indeed, when we conduct these tests we fail to detect any statistically significant results (Online Appendix Tables A7 and A8).

5.3. Correlation between beliefs and actions

We now turn to examining the correlation between beliefs and behavior. Our theoretical model predicts a positive correlation between players' beliefs on *r, E(r)* and *E(x)* and their choice of *x*. Moreover, Assumption (A) states that the correlation between the amount given and *E(r)* should be weaker in the High Variance treatment than in the Baseline and Low Average treatments.

We estimate these correlations in two ways. First, we exploit data from the B&A experiment, where we observe both beliefs and actions of the sample of dictators, to correlate each dictator's choice of *x* with his/her own beliefs *E(r)*, *E(x)* and *r*. Of course these are only correlations, since there is no exogenous source of variation within treatments. Second, we exploit both the A and the B&A experiments, following the approach of Krupka & Weber (2013). Namely, we estimate a conditional fixed-effects logistic regression by combining the beliefs elicited in Experiment B&A with data on dictators' actions from Experiment A. Panel data on actions and beliefs thus come from two different samples.

5.3.1 Action and beliefs in the B&A experiment

Table 5 reports the Spearman correlation coefficients between amount given and each of the three beliefs. In the first row we compute these correlations pooling Baseline, Low Average, and High Variance treatments together. The remaining rows consider each treatment in isolation.

The results displayed in Table 5 are consistent with the model's predictions: Column 2 reports the correlation between *x* and *E(r)*, which is positive when we pool all treatments together, as well as when we consider each of them separately. Moreover, such correlation is marginally statistically significant in the Baseline treatment ($p = 0.1005$) and significant in the Low Average treatment ($p = 0.0074$), while it is smaller in magnitude and far from statistically significant in the High Variance treatment. This finding, that in the High Variance treatment actions are less correlated with normative expectations, is consistent with assumption (A1) in the theoretical model.

Other beliefs are also, overall, significantly and positively correlated with donation amounts. In particular, in line with the findings of Bicchieri & Xiao (2009), the correlation between the amount given and empirical expectations, *E(x),* is always high and statistically significant. The correlation between *r* and the amount given by dictators is also positive and it is sometimes statistically significant.

Table 5. Correlation between actions and beliefs

	N	*E(r)*	*E(x)*	*r*
	(1)	(2)	(3)	(4)
All	290	0.1599	0.3262	0.1810
		(0.0063)	(0.0000)	(0.0020)
Baseline	96	0.1669	0.3282	0.1655
		(0.1005)	(0.0010)	(0.1035)
Low Average	98	0.2719	0.3508	0.2524
		(0.0074)	(0.0005)	(0.0131)
High Variance	96	0.0758	0.3164	0.1090
		(0.4631)	(0.0017)	(0.2904)

Note: each cell reports the Spearman correlation coefficient between the amount given by dictators in the B&A treatments and their own beliefs *E(r)* (Column 2), *E(x)* (Column 3) and *r* (Column 4). Baseline, Low Average, and High Variance treatments are pooled in the first row, and considered separately in the remaining rows. P-values in parentheses.

The pattern displayed in Table 5 is confirmed by regression analysis (Online Appendix Table A9), except that here normative expectations are not statistically significant or have the wrong (negative) sign. In a linear regression of the amount given on all three beliefs (for all treatments pooled together and for each treatment separately), positive expectations, *E(x)*, are always highly correlated with the amount given, consistently with Bicchieri & Xiao (2009). Moreover, subjective values, *r*, are more strongly correlated with the amount given in the Low Average Treatment, possibly suggesting an element of self-serving bias in the formation of normative standards (I behave consistently with my personal values when they are lower, as in the Low Average Treatment – see also Table 3 above); if so, however, self-serving bias seems to operate through personal values rather than through expectations of others' behavior.[31]

5.3.2 Matching actions in the A experiment with beliefs in the B&A experiment

Next, we estimate the correlations between actions and beliefs by conditional logit. For each treatment, the beliefs elicited in the B&A experiment reveal the perceived characteristic of each alternative course of action, in terms of personal values and normative and positive expectations. Note that here the beliefs associated to each action are the same for all individuals in the same treatment, but they possibly differ across treatments. We ask how these characteristics of beliefs elicited in experiment B&A correlate with the actions actually chosen by a different sample in experiment A (with the same treatment).

[31] The individual coefficients in Appendix Table A9 have to be interpreted with caution, since clearly the three beliefs indicators are highly mutually correlated (in the pooled sample, their pairwise correlation coefficients range between 0.36 and 0.54).

The binary dependent variable is an indicator of whether a certain action $a \in \{0,1,\ldots,10\}$ was taken in the treatments of Experiment A. The explanatory variables are different indicators of beliefs regarding action *a,* elicited in the treatments of Experiment B&A. The results, displayed in Table A.10 of the Appendix, confirm the patterns described above. When taken in isolation, beliefs (personal values, normative expectations and positive expectations) referring to a specific action are positively correlated with the corresponding action in all treatments. When all three beliefs indicators are included in the same regression, positive expectations are positively and significantly correlated with the corresponding action in two out of three treatments, again in line with the findings of Bicchieri & Xiao (2009), while the pattern of correlations for the other beliefs indicators is less precisely estimated. The appendix provides more detail.

6. Discussion

We have presented a new model of social norms consistent with some ideas in the literature on social norms, as well as experimental evidence testing that model. We have obtained two results. First, individuals who received the Low Average informational treatment give less as dictators, compared to individuals exposed to the default distribution. Second, individuals who received the High Variance treatment on average do not donate a different amount than the default group, but exhibit a higher variance in the amount given, as a group.

We also verified that the informational treatments change beliefs in the expected direction. We elicited three set of beliefs: 1. on the normative standards of others (i.e. what others regard as socially appropriate); 2. on positive expectations of others (i.e. what others will actually do); 3. on values (i.e. opinions on what is the right-thing-to-do, as opposed as to what others regard the right-thing-to-do). We used a novel elicitation method for 1 and 2, which allows us to observe the full distribution of normative and positive expectations over the entire range of possible actions. All the beliefs that we elicited are affected by informational treatments in the expected direction, although the effect is strongest on the beliefs over the normative standards of others. Namely, individuals exposed to the Low Average treatment perceive others as having the normative expectation of a lower amount given, while those exposed to the High Variance treatment have a more dispersed perceived distribution of the normative expectation of others. Moreover, these beliefs are correlated with actions in the expected manner. The correlation between actions and positive expectations is particularly robust, confirming previous findings by Bicchieri & Xiao (2009).

Overall, therefore, these results confirm that social norms influence actions through individual perceptions of the normative standards of others, and that such beliefs react to available information.

Moreover, the consensus around the social norm, measured by the variance of perceived normative beliefs, also matters. A more partial norm (i.e., one on which consensus is weaker) is associated with more dispersion in individual actions, suggesting that the social norm is less influential, consistently also with the ideas in Bicchieri (2017). However, less consensus around the contents of a social norm need not imply more selfish behavior, since a weaker social norm also gives more room to the influence of personal values.

Several important questions for future research stand out. First, we hope that our paper will stimulate theoretical efforts to model the impact of social norms in general games. Second, many follow-up experiments would seem interesting. For example, *why* do individuals react to the perceived normative standards of others? The literature on social identity suggests an answer: because individuals who identify with a social group behave consistently with their perceptions of how a typical member of the group ought to behave. This leads to the conjecture that increasing the salience of group identification, and creating situations of conflict between groups, would also increase the influence of group norms (defined as normative standards shared within the group). Results on cooperative behavior of individuals exposed to wars point in this direction (Bauer et al. 2016). Exploring the empirical validity of this conjecture, with particular reference to the social identity approach, is both feasible and interesting. More generally, studying the link between social norms and social identities, and how both are influenced by perceptions of group features and of group leaders is an important and promising area of research.

References

Allcott, H. (2011), “Social Norms and Energy Conservation.” *Journal of Public Economics* 95: 1082-1095.

Andreoni, J. & B. D. Bernheim (2009), “Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects.” *Econometrica* 77: 1607-1636.

Andreoni, J., Nikiforakis, N. & Siegenthaler, S. (2020), “Predicting Social Tipping and Norm Change in Controlled Experiments.” *NBER Working Paper 27310.*

Ayres, I., Raseman, S. & Shih, A. (2013), “Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage.” *Journal of Law, Economics and Organizations* 29(5): 992-1022.

Battigalli, P. & Dufwenberg, M. (2007), “Guilt in Games.” *American Economic Review: Papers & Proceedings* 97: 170-176.

Battigalli, P. & Dufwenberg, M. (2009), “Dynamic Psychological Games.” *Journal of Economic Theory* 144: 1-35.

Bauer, M., Blattman C., Chytilova, J., Henrich, J., Miguel, E. & Mitts, T. (2016), “Can War Foster Cooperation?,” *Journal of Economic Perspectives*, 30 (3).

Beaman, L., Duflo, E., Pande, R. & Topalova, P. (2012), “Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India.” *Science* 335: 582-586.

Bénabou, R. & Tirole, J. (2006), “Incentives and Prosocial Behavior.” *American Economic Review* 96: 1652-78.

Bernheim, B.D. (1994). “A Theory of Conformity.” *Journal of Political Economy* 102: 841-877.

Beshears, J., Choi, J.J., Laibson, D., Madrian, B.C. & Milkman, K.L. (2015), “The Effect of Providing Peer Information on Retirement Savings Decisions.” *The Journal of Finance* 70: 1161-1201.

Bicchieri, C. (2006), *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, Cambridge, MA.

Bicchieri C. (2017), *Norms in the Wild: How to Diagnose, Measure, and Change Social Norms*. Oxford University Press.

Bicchieri, C., Dimant, E. & Sonderegger, S. (2020), “It's Not A Lie if You Believe the Norm Does Not Apply: Conditional Norm-Following with Strategic Beliefs.” *SSRN Working Paper 3326146*.

Bicchieri, C. & Xiao, E. (2009), “Do the right thing: But only if others do so.” *Journal of Behavioral Decision Making*, 22(2): 191-208.

Binmore, K. (2010), “Social Norms or Social Preferences?” Mind & Society 9: 139-157.

Bursztyn, L., Gonzalez, A.L. & Yanagizawa-Drott, D. (2020a), “Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia.” *American Economic Review, forthcoming*.

Bursztyn, L., Egorov, G. & Fiorin, S. (2020b), “From Extreme to Mainstream: The Erosion of Social Norms Unravel.” *American Economic Review, forthcoming*.

Byrne, D.P., La Nauze, A. & Martin L.A. (2018), “Tell Me Something I Don’t Already Know: Informedness and the Impact of Information Programs.” *The Review of Economics and Statistics,* 100: 510-527.

Chong, A., Duryea, S. & La Ferrara, E. (2012), "Soap Operas and Fertility: Evidence from Brazil." *American Economic Journal: Applied Economics*, 4 (4): 1-31.

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.” *Journal of Personality and Social Psychology, 58*(6): 1015-1026.

d’Adda, G., Darai, D., Pavanini, N. & Weber, R.A. (2017), “Do Leaders Affect Ethical Conduct?” *Journal of the European Economic Association* 15: 1177-1213.

Danilov, A., Khalmetski K. & Sliwka D. (2018). “Norms and Guilt.” *CESifo Working Paper* 6999.

Drouvelis, M. & Nosenzo, D. (2013), “Group Identity and Leading-by-example.” *Journal of Economic Psychology* 39: 414-425.

Elster, J. (1989), “Social Norms and Economic Theory.” *The Journal of Economic Perspectives* 3(4): 99-117.

Ferraro, P.J., Miranda, J.J. & Price, M.K. (2011), “The Persistence of Treatment Effects with Norm-Based Policy Instruments: Evidence from a Randomized Environmental Policy Experiment.” *The American Economic Review* 101: 318-322.

Ferraro, P.J. & Price, M.K. (2013), “Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment.” *Review of Economics and Statistics* 95: 64-73.

Fehr, E. & Schurtenberger, I. (2018), “Normative Foundations of Human Cooperation.” *Nature* 2: 458-468.

Frey, B.S. & Meier, S. (2004), “Social Comparisons and Pro-social Behavior: Testing “Conditional Cooperation” in a Field Experiment.” *American Economic Review* 94: 1717-1722.

Geanakoplos, J., Pearce, D. & Stacchetti, E. (1989), “Psychological games and sequential rationality.” *Games and Economic Behavior* 1: 60-79.

Gerber, A.S. & Rogers, T. (2009), “Descriptive Social Norms and Motivation to Vote: Everybody’s Voting and So Should You.” *The Journal of Politics* 71: 178-191.

Gintis, H. (2010), “Social norms as choreography.” *Politics, Philosophy & Economics* 9, no. 3: 251-264.

Greiner, B. (2015), “Subject Pool Recruitment Procedures: Organizing Experiments with ORSEE.” *Journal of the Economic Science Association* 1(1): 114-125.

Jensen, R. & Oster, E. (2009), “The Power of TV: Cable Television and Women’s Status in India.” *The Quarterly Journal of Economics* 124: 1057-1094.

Krupka, E. L. & Weber, R. (2009), “The Focusing and Informational Effects of Norms on Pro-social Behavior.” *Journal of Economic Psychology* 30: 307-320.

Krupka, E. L. & Weber, R. (2013), “Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary?” *Journal of the European Economic Association*, 11(3): 495-524.

Landa, D. (2006), “Rational choices as social norms. ” *Journal of Theoretical Politics,* 18, no. 4: 434-453.

Lewis, D. (1969), *Conventions: A Philosophical Study*, Cambridge, MA: Harvard University Press.

López-Pérez, R. (2008), “Aversion to Norm-breaking: A Model.” *Games and Economic Behavior* 64: 237-267.

McDonald, R. I., Fielding, K. S., & Louis, W. R. (2013), "Energizing and De-Motivating Effects of Norm-Conflict." *Personality and Social Psychology Bulletin*, *39*(1), 57–72.

Michaeli, M. & D. Spiro (2015), “Norm Conformity across Societies.” *Journal of Public Economics* 132, 51-65.

Miller, D.T. & D.A. Prentice (2016), “Changing norms to change behavior. ” *Annual Review of Psychology* 67: 339-361.

Nikiforakis, N., Noussair, C.N. & Wilkening, T. (2012), "Normative Conflict and Feuds: The Limits of Self-enforcement." *Journal of Public Economics*, 96, 797-807.

Paluck, E.L. (2009), “What’s in a Norm? Sources and Processes of Norm Change.” *Journal of Personality and Social Psychology* 96(3): 594-600.

Peysakhovich, A. & Rand, D.G. (2015), “Habits of Virtue: Creating Norms of Cooperation and Defection in the Laboratory.” *Management Science* 62: 631-647.

Prentice, D. A. (2007), "Norms, Prescriptive and Descriptive." *Encyclopedia of Social Psychology*, 630-631.

Shang, J. & Croson, R. (2009), “A Field Experiment in Charitable Contribution: The Impact of Social Information on the Voluntary Provision of Public Goods.” *The Economic Journal* 119: 1422-1439.

Sugden, R. (1989), “Spontaneous Order, ” *Journal of Economic Perspectives* 3–4: 85–97.

Tankard, M.E. & Paluck, E.L. (2016), “Norm Perception as a Vehicle for Social Change.” *Social Issues and Policy Review*, 10: 181-211

Taylor, M. (1987). *The Possibility of Cooperation*, Cambridge: Cambridge University Press.