Measuring Norm Pluralism and Tolerance

Folco Panizza[a], Eugen Dimant[b,c], Erik O. Kimbrough[d], Alexander Vostroknutov[e]

[a] *IMT School for Advanced Studies Lucca*
[b] *University of Pennsylvania, Center for Social Norms and Behavioral Dynamics*
[c] *CESifo*
[d] *Smith Institute for Political Economy and Philosophy, Chapman University*
[e] *Maastricht University*

Abstract

This study introduces the Norm-Drawing Task, a novel approach to measure pluralism, or the coexistence of multiple normative beliefs in a given situation. By combining established methods, we identify heterogeneous normative beliefs in well-known economic games, challenging the typical assumption of a single prevailing norm. Moreover, we are able to link norm multiplicity to actual behavior. In a well-powered and pre-registered experiment, we observe that participants holding multiple normative beliefs are more tolerant and punish norm violations less frequently and less severely than those with a singular normative belief. This pattern suggests that norm multiplicity can lead to more lenient enforcement: pluralism and tolerance seem to co-exist. The implications of our study are broad, indicating that societal structures and policy decisions could be influenced by the underlying multiplicity of norms. Moreover, the Norm-Drawing Task, for which we provide a ready-made software implementation, offers a new avenue for exploring important societal issues like pluralistic ignorance and the dynamics of polarization.

Keywords: Distribution, Norm Elicitation, Social Norms, Norm Uncertainty, Tight and Loose Norms

JEL: C9, D01, D9

*This research was kindly funded by International Foundation for Research in Experimental Economics (grant no. 31-1-19); pre-registration: socialscienceregistry.org/trials/12012; software implementation: osf.io/7hc9v
Email addresses: folco.panizza@imtlucca.it (Folco Panizza), edimant@sas.upenn.edu (Eugen Dimant), ekimbrou@chapman.edu (Erik O. Kimbrough), a.vostroknutov@maastrichtuniversity.nl (Alexander Vostroknutov)

This version: November 30, 2023

> We must think in terms of degrees of normness, or the process of normative regulation, rather than in terms of norm as a thing. This makes it possible in research to investigate whether, in what form, and to what degree norms "exist," instead of taking them for granted.
>
> *Jay Jackson, 1966*

1. Introduction

There is a broad consensus among social scientists that norms permeate our lives, facilitating cooperation and coordination, distinguishing groups from one another, and shaping our beliefs and attitudes (see e.g., Sherif, 1936; Cialdini and Trost, 1998; Bicchieri, 2006; Henrich, 2017). Norms are conceptualized as mutually reinforcing patterns of beliefs and behavior, and they are often divided into an injunctive aspect, which refers to what ought to be done, and a descriptive aspect, which describes what is actually done by people in a group. These two elements of a norm are essential because the most popular accounts of norm-driven behavior assume that people are motivated to conform to norms that they believe others will follow and that they believe others expect them to follow in kind. In these accounts, changes to either aspect of beliefs can have important consequences for behavior.

To better understand the role of norms in shaping behavior, social scientists have thus recently worked to develop methods for measuring them in all their aspects. While measuring behavior is straightforward, measuring beliefs is more difficult — all the more so when these beliefs are about non-verifiable objects like normative judgments regarding what ought to be done. Nevertheless, techniques for measuring shared normative beliefs have been fruitfully employed to improve our understanding of individual decision-making, strategic decision-making, and political expression and behavior, to name just a few examples (Krupka and Weber, 2013; Bicchieri and Xiao, 2009; Barr et al., 2018; Dimant, 2019; Bicchieri et al., 2022; Pickup et al., 2021, 2023).

In some cases, such as the widely studied Dictator Game, in which one person decides how to divide an unearned sum of money with another person, average normative beliefs have been shown to be strikingly consistent across a wide range of subject pools and countries (Krupka and Weber, 2013; Kimbrough and Vostroknutov, 2016, 2018). However, a key implicit assumption of existing methods for measuring norms is that there is only one norm. Usually, the question is "What is the norm that people adhere to in a certain environment?" While this is a legitimate question, in many cases the implicit assumption that only one norm exists may obfuscate the reality that there is *norm multiplicity* or many normative beliefs in the same environment that can possibly contradict each other.

One case in which there may be multiple norms is when those norms are tied to distinctive identity groups (Groenendyk et al., 2023). For example, during Covid-19 beliefs about masking were closely related to political identity. Some people believed that the norm was

to wear a mask in public places, whereas others thought that the norm was to keep the pre-Covid status quo, not wearing a mask. This is a typical example of norm multiplicity where it was clear to everyone that there were two norms followed by different groups of people. This situation created a lot of confusion and normative disagreement and did not promulgate healthy behaviors (Gelfand et al., 2021a,b; Dimant et al., 2022a).

However, even in the abstract environment of the Dictator Game, questions have been raised about whether the stable *average* pattern of normative beliefs masks several underlying heterogeneous types (Kimbrough et al., 2022).[1] And both theory and evidence have been accumulating that multiple norms can coexist even in the absence of sharp identity boundaries, that events can trigger transitions between them, and that their co-existence can have important consequences for behavior (e.g., Centola et al., 2018; Fromell et al., 2021; Dimant and Gesche, 2023; Dimant et al., 2023; Merguei et al., 2022). These consequences can be dire: some have argued that norm multiplicity can create normative disagreements or deterioration of norms that can hurt cooperation or even lead to a conflict or war. But they can also be salutary: mutual recognition of differences of viewpoint can breed tolerance.

Yet, as noted above, current methods are not well-suited to identifying mulitiplicity. To see the problem, consider the two main approaches to measuring norms: methods that elicit point estimates of normative beliefs (e.g., Bicchieri and Xiao, 2009; Krupka and Weber, 2013); and methods that elicit distributions of beliefs (e.g., Dimant et al., 2022b; Dimant, 2023a). The first approach, by design, elicits only one normative view and thus can only provide "average" normative beliefs in situations with multiple norms.[2] The second approach can, in principle, reveal that multiplicity exists, e.g. if subjects report polarized beliefs about the appropriateness of an action, but eliciting the distribution does not allow one to see the underlying types out of which the distribution is composed. For example, in a two-action choice set, if beliefs are polarized for both actions, the researcher can say that there seem to be multiple norms, but the same distribution of beliefs could be generated by a variety of underlying norm types. The key conceptual issue is that a norm is not just characterized by the most appropriate action but rather a set of beliefs about the *relative* appropriateness of all possible actions. Without knowing for each point in the distribution of beliefs about one action how it is connected to beliefs about other available actions, a method that measures distributions cannot reveal the nature of heterogeneity.

[1] They use latent class models to identify 5 underlying normative belief types of which the average pattern in the Dictator Game is composed; they find that though the average normative belief is strikingly similar to that seen in other settings, there is some evidence of heterogeneity in the underlying composition of types across two distinct cultural settings. As the authors point out, given that norms were elicited using the coordination game method due to Krupka and Weber, the fact that subjects report heterogeneous beliefs in their study can only arise if either subjects make mistakes or there is genuine uncertainty about which of many beliefs is actually most common in the population.

[2] The exception here is when norms are known to be linked to distinct reference groups; then, by eliciting norms separately for each reference group, multiplicity is observable.

In light of these issues, it is essential to 1) develop methods for identifying norm multiplicity, in order to fully understand the normative landscape against which individuals evaluate and adjust their social behavior, and 2) begin to augment theories of norm-driven behavior to explicitly incorporate the effects of norm multiplicity on decision-making. In this paper, we make progress towards both of these goals.

First, we propose a new Norm-Drawing Task that allows us to explicitly elicit multiple normative views in choice problems. In this two-step procedure, we first elicit a collection of normative beliefs (instead of one) using an interface inspired by Crosetto and De Haan (2023), in which subjects trace paths through the action space assigning a relative appropriateness to each action, and then we apply the Belief Elicitation by Superimposition Approach or BESA (Fragiadakis et al., 2019) to the resulting collection of normative views to reveal participants' beliefs about how often the chosen views occur in the population (or experimental session). The combination of the two procedures generates an incentive-compatible method that not only elicits multiple norms in a given situation but also lets participants determine endogenously how many views they think there are.

The Norm-Drawing Task generates a new type of data in which we know for each participant what norms they believe apply in a given situation. To illustrate how these beliefs can influence behavior, we propose a simple theoretical argument built on the ideas developed in Kimbrough and Vostroknutov (2023b,a) and Merguei et al. (2022) that relate the multiplicity of norms to the punishment of norm violators. Specifically, we argue that norm multiplicity should make the punishment of norm violations smaller and less frequent (as compared to the case of singe-norm beliefs) given that, in multiple norms environments, it is unclear the violations of which norms should be punished (in which case the decision-maker chooses the "cheapest" norm for punishment), or whether the decision-maker considers one "expected" norm that also generates less punishment. As noted above, this implied reduction in punishment can be seen as a mechanism by which pluralism breeds tolerance.

In the experiment, we test the Norm-Drawing Task by eliciting beliefs in a 2×2 design with one dimension corresponding to the type of game (a Dictator Game or a three-player Allocation Game similar to those in Engelmann and Strobel, 2004) and another dimension corresponding to presence or absence of a passive charity player. With the first dimension, we check whether games with no equality-efficiency tradeoff (the Dictator Game) differ in the number and nature of normative views from games that contain an equality-efficiency tradeoff (the Allocation Game). With the second dimension, we test if replacing one human participant with a charity changes the degree of multiplicity of norms, on the assumption that a charity will be seen as more deserving of resources (we assume) than will another survey respondent. At the same time, we collect data on punishment choices within the same participants to connect their normative beliefs to their behavior and to test if beliefs about norm multiplicity do, in fact, reduce punishment.

We find that the punishment patterns indeed fit our theory: participants who express a single-norm belief punish significantly more harshly and also more often than those who express multi-norm beliefs. With regard to the extent of multiplicity, we find that in both the Dictator Game and the Allocation Game participants report around 4 to 5 normative views on average. The evidence of multiplicity in the dictator game may be surprising, but it provides additional support to the findings from Kimbrough et al. (2022) noted above. The fact that participants recognize multiple normative views in such a simple game as the Dictator Game suggests that we should expect norm multiplicity to be ubiquitous in real environments and that it is important to take its potential consequences into account.

Our results can have far-reaching implications. For example, Gelfand et al. (2011) observe that human societies can be divided into two general classes: tight and loose societies that are different in terms of harshness of punishment of norm violations. This has consequences for the quality of institutions, welfare, and many other important economic indicators (Gelfand et al., 2024). We suggest that it might be fruitful to conceptualize tightness and looseness in terms of norm multiplicity: tight societies have a single, widely shared norm with correspondingly harsh enforcement, which leads to a more structured society with orderly rules of behavior; loose societies have a multiplicity of norms, which leads to less enforcement and more tolerance of difference. However, in a dynamic world such looseness may also lead to the gradual deterioration of norms (Henrich, 2017), and more chaotic institutions. We suggest that our method could be used as a complement to existing measures of tightness and looseness to help further understand the mechanisms that link this cultural dimension to behavior.

The Norm-Drawing Task can also be used for other purposes. For example, detecting pluralistic ignorance (Bicchieri, 2006) or polarization (Iyengar et al., 2019; Bursztyn et al., 2020; Levy, 2021; Dimant, 2023b). More broadly, the task allows us to reconstruct the normative landscape that participants believe they live in: the number and kind of norms that they express in certain contexts can signal what sort of attitudes and heterogeneity in beliefs they expect to deal with in reality and what consequences for behavior this might have.

The paper is structured as follows. In Section 2 we review the existing norm elicitation techniques and argue why they cannot adequately capture norm multiplicity. In Section 3, we present the Norm-Drawing Task and propose the theoretical argument about the data it generates. In Section 4, we describe our experimental treatments and formulate the hypotheses. In Section 5, we report the experimental findings. Section 6 interprets our findings, proposes the future directions of research, and concludes.

2. Existing Norm Elicitation Techniques

Several approaches to eliciting normative beliefs exist in the literature that vary depending on the type of norm under investigation (Charness et al., 2021). For instance, one of the straightforward methods to elicit descriptive norms (where the appropriateness of outcomes depends on the observed behavior) is to use the behavior itself to incentivize truthful revelation of beliefs. Bicchieri and Xiao (2009) elicit descriptive norms using a frequency method that incentivizes participants to guess objectively verifiable information about the behavior of others in the experiment. After choosing how to distribute payoffs in a Dictator Game, participants are asked to predict the number of others who split the money approximately equally (this information is available to the experimenters from the first part of the experiment, and subjects are paid a fixed sum if they guess correctly). Norm elicitation of this kind can also rely on different mechanisms, such as scoring rules or the interval method, as long as the underlying behavior—that serves as a benchmark for incentivization—can be objectively measured.

By contrast, injunctive normative beliefs refer to objects that cannot be independently and objectively verified. Thus eliciting these beliefs is trickier. One approach, due to Bicchieri and Xiao (2009), adopts a two-step process for eliciting beliefs that combines a non-incentivized self-report of what the subject personally believes is normatively best and the frequency method described above, in which subjects are incentivized to guess the personal beliefs reported by others. For the dictator game, participants are asked whether dictators should split the money equally, and then they are asked how many other participants answered "Yes" to the first question. This approach works if subjects report their personal beliefs honestly, but given that personal beliefs are internally held and unverifiable, there is no assurance that respondents would reply honestly, especially concerning sensitive topics.

Another elicitation method that does not rely on objectively verifiable information was introduced by Krupka and Weber (2013). This method elicits social norms by incentivizing subjects to report shared beliefs via coordination games. Specifically, participants are asked to predict how others would rate an action in terms of its social appropriateness (e.g., inappropriate, neutral, appropriate). To win a prize, participants must match their guess about the appropriateness rating to the most common guess made by others. They are paid only if their guess matches the modal guess. This method works if, to solve the coordination problem, subjects rely on existing shared beliefs about what is socially (in)appropriate as a focal point.

While this method is incentive-compatible, it relies on an implicit assumption that there is only one norm in the environment under consideration. This assumption can be justified in certain cases where there is clearly only one norm (e.g., stealing is bad). However in general and even in seemingly simple experimental tasks like the Dictator Game as we show below, it is not obvious that only one norm exists. Thus, the coordination-game methods

derived from Krupka and Weber (2013) may not yield precise norm estimates in scenarios where it is difficult to coordinate on the appropriateness of actions. This is particularly relevant for contexts where multiple norms are present. In fact, both the two-step and the coordination-game methods focus on the most frequent evaluation, ignoring any information about the multiplicity of norms or their uncertainty.

The approaches described above are justified when researchers are interested in estimating only the average social appropriateness or in contexts where only a single norm is clearly focal. As we argue, however, it is not obvious to expect a single norm in many environments, nor that norms will remain unchanged, for instance, because minority views take hold. Even in contexts where apparently only one norm exists, diverging perspectives might lurk beneath the surface and influence behavior.

One approach to dealing with these concerns has been to consider that normative beliefs may be held with uncertainty. Thus, some methods have been introduced to elicit not just point estimates of normative beliefs but also estimates of the distribution. One such approach is the quadratic scoring rule (QSR) that is frequently used to incentivize the elicitation of various kinds of beliefs (Artinger et al., 2010). With regard to norms, Dimant et al. (2022b); Dimant (2023a) uses QSR to estimate beliefs about descriptive norms in a one-shot Public Good Game. Participants are paired with each other before the game and have to guess how the opponent will play by assigning probabilities to each possible choice of contribution. Then, the actual contribution of the opponent is revealed, and the participants are paid based on the probability they have assigned to the true choice according to QSR.

As with any scoring rule method, QSR works best when predictions are compared to an objective benchmark, which makes it a suitable elicitation method for descriptive norms (for which the objective benchmark exists in the form of observed behavior). However, when the benchmark is not objectively verifiable—which is the case with the injunctive norms—QSR does not provide an incentive-compatible mechanism.[3] Such a situation may not be desirable. In such cases, other methods are used that overcome this problem at the expense of simplicity.

Another approach employs a variant of the Colonel Blotto game to elicit beliefs (Peeters and Wolk, 2019). The Blotto game was first formalized by Borel (1921) and developed in Roberson (2006). Here, each option is conceived as a "battlefield" where players place their "forces" (i.e., tokens). Players try to predict which battlefield will have the largest number of forces deployed since a fixed prize will be assigned to participants according to the proportion of tokens on that battlefield (the more tokens the higher the probability of the prize). The best strategy in this game is to weigh each battlefield according to its probability of being the one with the most forces and to place one's tokens on that battlefield.

[3]What's more, QSR is only incentive compatible under risk neutrality (Offerman et al., 2009).

Colonel Blotto could easily be adapted to elicit distributions of normative beliefs by defining each appropriateness rating as a battlefield. While this approach allows for eliciting a unique norm, its winner-take-all incentive structure risks concealing the presence of multiple norms. As an example, think of an action that is considered very appropriate by a majority of the population and very inappropriate by a minority. Since in the Colonel Blotto game, only one battlefield will be chosen for payment, the best strategy for participants is to focus on how most players will place their tokens. This makes participants think about the norm endorsed by the majority, and disregard the tokens placed by the minority. The elicitation procedure thus risks not reflecting participants' actual distribution of normative beliefs by not giving them incentives to reveal their beliefs about potential minorities.

Another method for eliciting uncertain beliefs is the Belief Elicitation by Superimposition Approach or BESA (Fragiadakis et al., 2019). This is a scoring-rule method that elicits belief distributions while preserving incentive compatibility. The procedure for BESA is the following: subjects place a number of tokens (conventionally 100) over the set of possible responses. This creates a histogram representing their belief about the distribution of responses. Fragiadakis et al. (2019) use the method to elicit empirical beliefs about observable behavior; in their design, the more closely the histogram matches the empirical distribution of behavior, the higher the probability that the subject wins a fixed sum. The probability of payment is proportional to the area of superimposition of the own token distribution (histogram) and the average distribution. If for instance, the participant exactly guessed the distribution, the probability of payment is 1. The more distant the participant's guess is from reality, the lower the probability of getting paid.

Aycinena et al. (2022) show how to adapt this method to measure uncertain normative beliefs by extending the two-step method developed by Bicchieri and Xiao (2009); in their design the histogram produced by subjects is compared to the empirical distribution of personal normative beliefs reported in step 1.

Here we note that a further extension of this approach can make it analogous to the coordination game method introduced by Krupka and Weber (2013). In this variant of BESA, subjects are instructed to assign tokens across the appropriateness ratings for each action according to how they believe others will place their own tokens. Matching others' guesses about the distribution increases the chances of payment: once all players have placed their tokens, an average distribution is computed across all participants. The closer the participant's distribution is to the average, the higher the likelihood of winning a fixed sum of money.

This variant of BESA can be used to estimate the distribution of beliefs over single actions. Consider blowing one's nose in public: respondents may be asked to estimate the share of individuals who think that nose-blowing is appropriate, and the share of those who think is inappropriate, then to estimate the share who think it is appropriate to sniff, and so

on. However, since beliefs estimated are over single actions, it is not possible to determine the share of individuals who think, for instance, that sneezing is inappropriate and sniffing is appropriate. That is, BESA cannot tell us which norm functions (mappings from actions to relative social appropriateness) are being mixed, since the distributions just sum them all together.

In sum, while the elicitation of distributions of normative beliefs using QSR, Colonel Blotto games, or BESA contributes to the detection of variability and uncertainty in norm perception, these approaches do not allow us to disentangle the nature of that uncertainty and do not reveal the multiple norms that might lead to diverging behavioral outcomes. Thus, we propose a method that incentive compatibly addresses these concerns.

3. The Norm-Drawing Task

In this study, we present and test a new incentive-compatible method to elicit a multiplicity of normative beliefs—such as social norms—in a given population. This method aims to identify and disentangle different normative beliefs that might exist in parallel and to identify how important each possible norm is relative to others.

The elicitation method follows a two-step procedure: first participants define the set of beliefs that exist in the scenario at hand, then they guess how these beliefs are distributed in the population. We will describe the second step first for illustrative purposes.

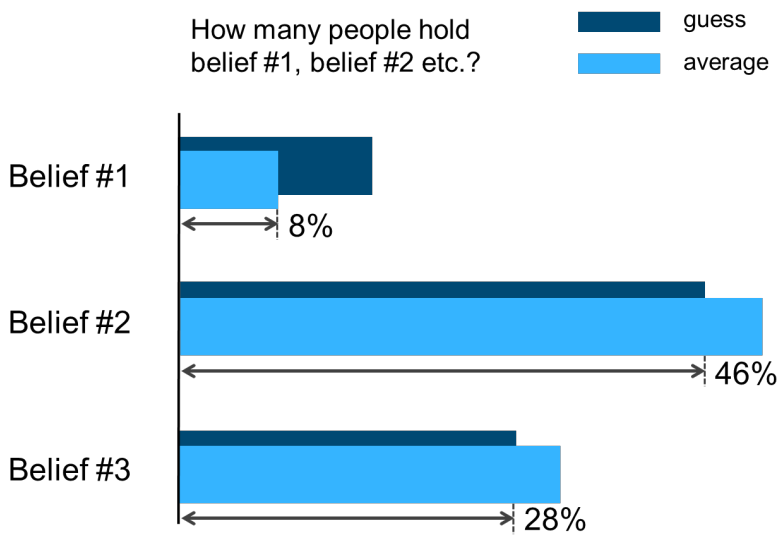

Figure 1: An example of token distribution. Participants make their guesses about the set of beliefs, then their distribution is compared to the average distribution made by other participants. Their probability of winning the prize for this task is equal to the sum of the minimum between their guess and the average guess for each belief. In this example, the sum (and probability of winning) equals to 82%.

To guess the frequency of different normative beliefs in the population, participants allocate 100 "people" (i.e., tokens) between different options that correspond to different

normative beliefs (or norms) chosen in the first step. For example, if there are two norms about the appropriate way to finish a meal (emptying the plate vs. leaving some food), a participant has to guess how many people follow each of the two norms. To make this procedure incentive-compatible we use BESA described above (Fragiadakis et al., 2019). After participants allocate tokens among the options, their allocation is superimposed on the average distribution. The probability of winning a prize is proportional to the overlap between the two (Figure 1). For n norms, we compute the overlap as $\sum_{i=1}^{n} \min\{p_i, E[p_i]\}$, where p_i is the number of people placed in norm i and $E[p_i]$ is the average number of people assigned to norm i by all participants.

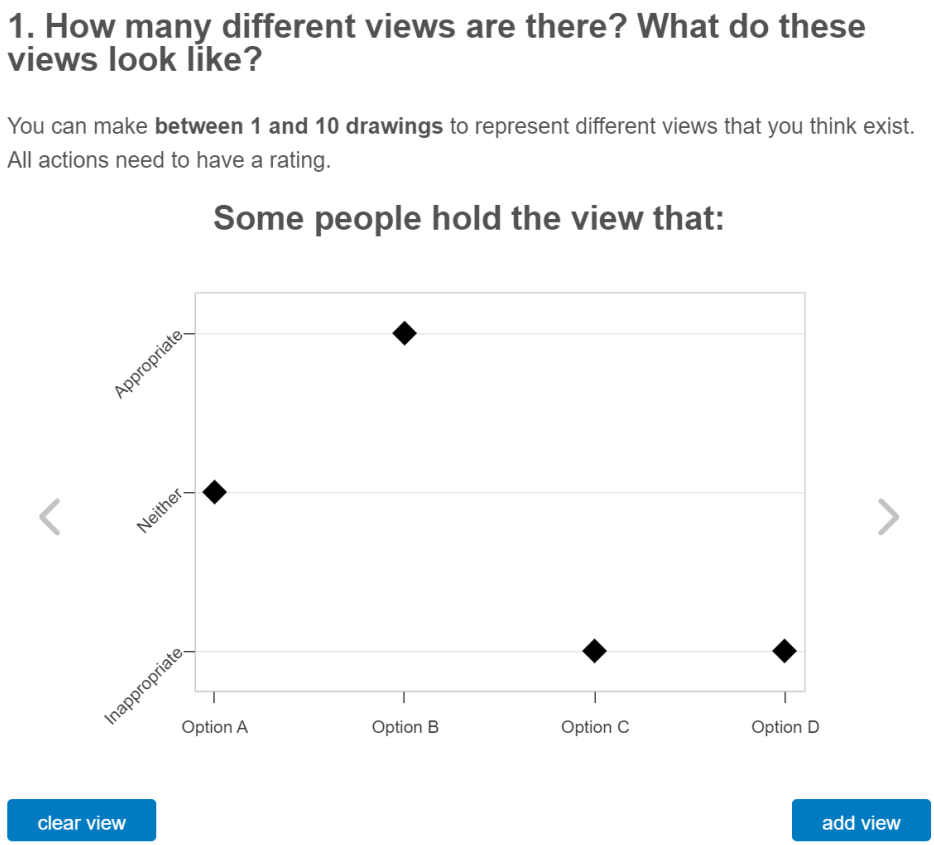

Figure 2: Graphical interface for entering ratings for each action. Participants start with a blank drawing and can click or drag the dots to rate the appropriateness of each action. Participants can also navigate through the different drawings, delete drawings, or add up to ten different drawings. Actions are connected by lines if on a continuous scale, otherwise unconnected.

Which beliefs are included in the second step is determined in the first step. Here, participants guess the beliefs that exist in the population. Participants do this by "drawing" the function corresponding to each belief (see Figure 2). Participants rate each action's appropriateness on a 3-Likert scale (appropriate, neither appropriate nor inappropriate,

inappropriate).[4] As an example, consider the meal scenario mentioned above: one norm might be that emptying the plate is appropriate while leaving leftovers is inappropriate, and another norm might correspond to the situation where both actions have the same appropriateness. Participants can guess from one to up to ten distinct norms through the graphical interface. Crucially, the BESA incentive scheme encourages participants to draw exactly the beliefs that they think exist in the population, no more and no less. This is because both drawing a belief that no one else has drawn or failing to draw a belief that others have drawn can only reduce the likelihood of winning in step two. Thus, our norm-drawing task is incentive-compatible and allows us to elicit the shape of each individual norm function (the function maps all available actions into social appropriateness) as well as the frequency of this norm function in the population.

Given these new data that were not available in the previous elicitation methods, we propose a brief theoretical argument on what can be expected in terms of behavior in our experiment when participants perceive a multiplicity of norms as compared to only one norm. Suppose that we elicit normative beliefs about some set of actions A (e.g., $A = \{\text{empty plate}, \text{leave food}\}$). Then in the first step, our task produces n norms represented by functions $\eta_i : A \to R$ for $i \in [1, .., n]$ where $R = \{\text{inappropriate, neutral, appropriate}\}$ is the scale of appropriateness chosen for the experiment. In the second step, participants choose numbers $p_i \in [0, 1]$ that represent the proportions of the population that use norm η_i. We impose the condition $\sum_{i=1}^{n} p_i = 1$.

With these preliminaries, we can now think about the behavioral implications of the perceived multiplicity of norms. Specifically, we consider the punishment of norm violations that can be meted out by a participant who thinks there are many norms that apply to a given situation as compared to a participant who thinks there is only one norm. To do that, we follow Kimbrough and Vostroknutov (2023a) and suggest that the amount of punishment for violating some norm η_i is proportional to the resentment that the chosen action evokes. We model the resentment from choosing action $a \in A$ as the difference in social appropriateness between the action $a_i^* \in A$ that gives η_i-maximal appropriateness ("the norm" or the most appropriate action according to η_i) and the appropriateness of a. Thus, the resentment at a given η_i is given by $r(a|\eta_i) = \eta_i(a_i^*) - \eta_i(a)$. This number can be interpreted as the severity of norm violation when a was chosen given that a_i^* was available.

Now, we compare the resentments coming from a participant who believes that there are many norms with a participant who believes there is only one. To model the participant

[4] We ran a series of calibration pilots and they suggested that a higher number of options would reduce participants' ability to coordinate in the specific scenarios tested. Note that the number of options can be adapted depending on the scenario and research questions.

with many norms, we follow Merguei et al. (2022) and consider an "expected" norm function

$$\mu(a) = \sum_{i=1}^{n} p_i \eta_i(a)$$

for all $a \in A$. The norm function μ can be thought of as the expected norm function that evokes resentment in the participant who perceives norms η_i with $i = 1, .., n$. The question now is Does μ evoke more or less resentment than individual η_i? To answer this, suppose that $m^* \in A$ maximizes μ, or that m^* is the most appropriate action according to μ. Then

$$r(a|\mu) = \sum_{i=1}^{n} p_i(\eta_i(m^*) - \eta_i(a)) \leq \sum_{i=1}^{n} p_i r(a|\eta_i).$$

The last inequality follows from the fact that m^* is in general not the same as individual a_i^* that maximizes η_i. Moreover, this inequality implies that the expected resentment after a given μ (from the participant who perceives multiple norms) is no larger than the expected resentment after a in the population consisting of proportions p_i of agents who follow explicitly η_i. This already demonstrates that we should expect less punishment on average from participants with multiple norms than from participants who only believe in one norm.

An additional argument that participants with multiple norms punish less than participants with one norm comes from the results of Merguei et al. (2022). The authors show experimentally that their participants, when facing two norms η_1 and η_2 in the Dictator Game after observing an offer a, choose to punish using the norm (η_1 or η_2) that prescribes cheaper punishment (generates less resentment $r(a|\eta_i)$). Thus, it is also possible that the participants with multiple norms compute resentment as

$$r(a|\mu) = \min_{i=1..n} r(a|\eta_i).$$

In this case, the punishment that they should choose will be (weakly) smaller than all punishments prescribed by individual norms η_i.

We can conclude that theoretically, we should expect less punishment from the participants with multiple norms than from those with only one. We test this conjecture below.

4. Methods

4.1. Experimental Treatments

We use the norm-drawing task in two scenarios: a Dictator Game (DG) and an Allocation Game (AG) adapted from Engelmann and Strobel (2004). In the Dictator Game scenario, the dictator splits a pie of size 4 (experimental currency points; 1 point = \$0.50) with another unknown person. There are five possible allocations, from keeping all four points to giving all four points. In the Allocation Game scenario, a person has to choose between

four possible allocations of money among themselves and two other people: a selfish option, an equitable option, a maximin option, and an efficient option (see Table 1).

Table 1: Distribution of experimental currency in the Allocation Game.

Allocation	selfish	equitable	maximin	efficient
Person A (chooser)	6	2	4	5
Person B/charity	0	2	3	5
Person C	0	2	3	1
sum of payoffs	6	6	10	11

We chose to test our task in DG and AG for the following reasons. First, both of them have been extensively studied in experimental economics and we have a rather good idea about what to expect in terms of behavior. We also know from previous experiments (e.g., Kimbrough and Vostroknutov, 2018) how norms in DG, as measured in the task by Krupka and Weber, should look like. This presents us with a partial benchmark to compare our results with since to our knowledge no one previously measured normative beliefs in AG. Second, the two games differ in terms of the number of intuitively plausible norms that we found in the literature. While in the DG there is a broad consensus that equal split is the focal, most appropriate outcome, the AG is notorious for generating heterogeneous behavior where some people favor efficiency and some maximin norms (Engelmann and Strobel, 2004; Baader and Vostroknutov, 2017). Thus, we want to test our task in these games to see if the intuitive presence of multiple norms in AG and a single norm in DG translates into similar observations in the norm-drawing task.

As an additional test of the norm-drawing task, we examine whether responses in the two scenarios depend on who the recipients are. For this reason, we include an alternative version of the two games with one recipient being replaced by a charity. The charity organization, Hellen Keller International, was selected based on the evaluations by GiveWell, a non-profit that helps donors figure out how to maximize their impact. In addition, Hellen Keller operates both in the U.S. and worldwide, so it can appeal to both conservative and liberal-leaning individuals (Pizziol et al., 2023). In the charity version of the Dictator Game, the dictator chooses how to distribute currency between themself and the charity. In the Allocation Game, the decision-maker allocates points between themself, the charity, and a second person (Person C in Table 1).

4.2. Experimental Design

In the experiment, participants complete one of the four treatments: 2 scenarios (DG or AG) $\times$ 2 types of recipients (standard or charity). The full outline of the four questionnaires is available on the online repository for this experiment (osf.io/yh6gd).

In all experimental treatments, participants first complete the norm-drawing task for the

specific scenario (Dictator or Allocation Game). The prize is set to 6 experimental currency points. After the norm-drawing task, participants complete a third-party punishment game on the same scenario: participants are endowed with 4 points (in the DG scenario) or 6 points (in the AG scenario) and can choose to spend any amount to reduce the earnings of the dictator/chooser by an equal amount. Participants respond using a strategy method: they decide how much if at all, to sanction each possible action in the game described in the scenario. They are then paired with an actual player in the game, and their decision is implemented based on that player's action. Participants could decide to sanction more than the player's earnings (e.g., 4 points even if the player only gets 2 points); in this case, the player received 0 points but was informed about the size of the sanction.

In the third part of the experiment, participants acted as dictators/choosers to provide a match for other participants who were choosing punishment. However, in order to prevent the previous tasks from influencing their responses in the game, participants played the game of the other scenario, keeping the presence or absence of the charity fixed. For example, participants, who drew norms in the standard DG scenario, made choices in the standard AG, and vice versa. Participants were aware of the payment scheme, and thus knew that their actions could be sanctioned.

4.3. Hypotheses

The experimental design, hypotheses, and analyses were pre-registered on the website of the American Economic Association (RCT ID AEARCTR-0012012). Any amendment to the original protocol is presented next to the related analysis with an explanation[5]. Multiple comparisons are corrected using the false discovery rate method (Benjamini and Hochberg, 1995), and null results are not interpreted unless appropriate methods are available (e.g. Bayes Factor, equivalence testing). Square brackets indicate 95% confidence intervals. All tests are conducted in R (R Core Team, 2022). We formulate four hypotheses.

H1: non-randomness of responses. In the norm-drawing task, participants will coordinate on one or multiple norms in a non-random manner: that is, the total number of norms guessed and the agreement between participants cannot be explained by participants simply drawing beliefs at random. Assuming the null hypothesis that participants draw beliefs at random, the relative frequency of each possible norm will be n/N, where n is the average number of norms guessed by participants, ranging between 1 and the maximum possible (i.e., 10), and $N = r^a$ (i.e., all possible norms that can be guessed), where r denotes the number of possible ratings (e.g., appropriate/neither/inappropriate, $r = 3$), and a denotes the number of possible actions (5 in the dictator game, 4 in the allocation game). To test

[5]The original Hypothesis 4, comparison with findings in Engelmann and Strobel (2004), is presented in Appendix A.1 as it is an auxiliary hypothesis that does not contribute to testing the main theoretical predictions of the method.

for randomness, we compare observed frequencies to this discrete uniform distribution using a chi-squared test. We repeat the test for each experimental treatment.

H2: norm multiplicity. In the norm-drawing task, some participants will express multiple norms and some will express only one norm. This will be the result of a mixture of H2A) participants acknowledging only one norm (e.g., the one they follow, as in the case of false consensus) and H2B) participants also considering other norms that differ in terms of what is most appropriate. To test for the existence of these different kinds of beliefs we will look at descriptive statistics. For H2A we will count the number of participants who guess a single norm; for H2B we will count the number of participants who guess multiple norms with mutually exclusive most appropriate actions.

H3: more punishment with unique norms. Participants who guess the existence of only one norm with a single most appropriate action (i.e., who do not consider norms with multiple most appropriate ratings) will punish non-appropriate actions (actions rated "inappropriate" or "neither appropriate nor inappropriate") more frequently (H3A) and with a higher punishment than all other participants (H3B). These predictions come from the theoretical argument at the end of Section 3. To test H3A, we use a mixed-effects logistic regression where the dependent variable is binary punishment (1 if action is punished, 0 otherwise), and a dummy variable indicating whether the participant guessed a single norm with a single action rated as most appropriate (1 if yes, 0 otherwise) as an independent variable. Participant ID is included as a random intercept. To test H3B, we use a mixed-effects linear regression where the dependent variable is punishment (from 0 to maximum). All other variables are the same as above. As an amendment to the original hypothesis, we include two dummy variables for the scenario and the presence of the charity, as these two variables could also contribute to punishment rates. The results are robust to whether or not these variables are included. As a second amendment, we excluded punishment rating and its interaction with the dummy as independent variables, as we were interested in the average effect of punishment across ratings. In Appendix A.2 we report the results of the original pre-registered test, which yields the same results.

H4: charity as coordination device. The number of guessed norms will, on average, be smaller in the charity scenarios than in the standard scenarios. This should be the case because the norms concerned with giving to charities are more straightforward (one should give to charities) than the norms related to humans. We will run a Poisson regression with the number of guessed norms as the dependent variable and scenario and charity presence as predictors. We will then test whether, for each scenario, the number of norms is higher in versions that do not include a charity as an agent (i.e., we expect more variability).

4.4. Sample Size

We sought to recruit 800 participants, 200 per experimental treatment. Although the sample size was computed based on budget constraints, we estimated the minimum detectable effect size for our main hypothesis H1. Assuming that all participants guess the maximum possible number of norms (10 in the experiment), given all possible norms that can be guessed in the scenario with the most actions (N = 243), and a sample of 200 participants per test, then our MDE would be $w = 0.20$ with $\alpha = 5\%$ and a power $(1 - \beta)$ of 95%.

We ended up recruiting 820 U.S. residents on Prolific, a survey platform specialized in online experiments. The mean age was 42 ($SD = 14$), 49% were female, 49% were male, and 2% indicated other or preferred not to state their gender. Participants were balanced across treatments (standard dictator game $N = 197$, charity dictator game $N = 226$, standard allocation game $N = 202$, charity allocation game $N = 195$) and so were age and gender.

5. Results

5.1. Descriptive Statistics

Figures 3 and 4 plot histograms of the number of norms drawn in the norm-drawing task. Participants drew on average 4 to 5 norms ($SD = 3$). This suggests that the multiplicity of norms is a rather common phenomenon even when we consider very standard, typical allocation tasks like the Dictator and Allocation Games.

Figures B.6 to B.9 (see Appendix B) present the five most weighted norms and the weighted average of all norms in each treatment. Notably, while some norms were given considerable weight, the average number of tokens placed on any given norm was at most 41/100 (charity allocation game); so according to most participants, no single norm is followed by the majority of people.

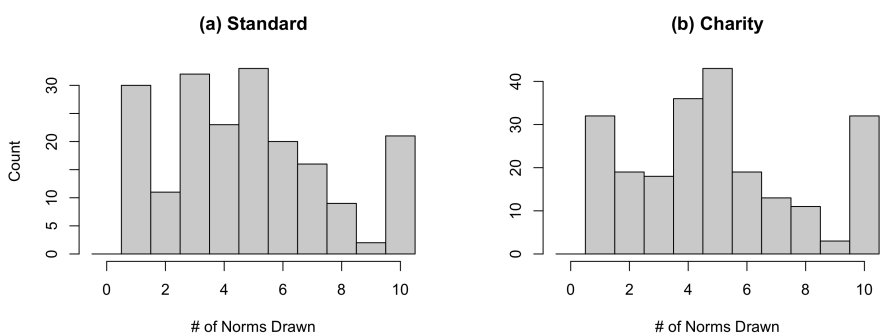

Figure 3: Histogram of Number of Norms Drawn by subject, DG.

These data suggest that traditional norm elicitation procedures like the standard Krupka-Weber task would have concealed the variance in responses. It is interesting to mention

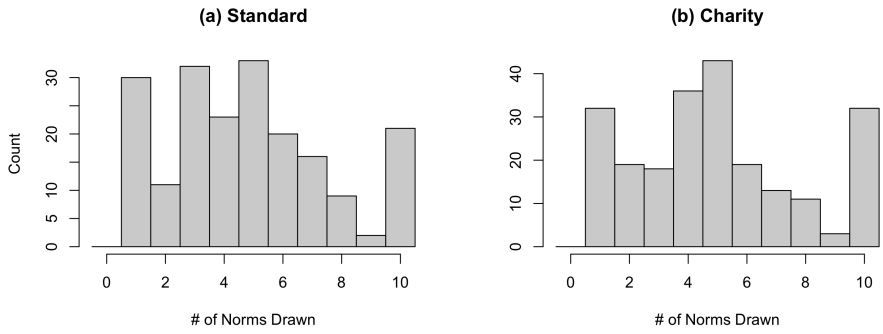

Figure 4: Histogram of Number of Norms Drawn by subject, AG.

though that the standard Krupka-Weber task seems to work in the sense that it produces average normative ratings in the DG that are very similar to our aggregates (compare, for example, Figure 2 in Kimbrough and Vostroknutov, 2018, to the bottom right graph on Figure B.6). Thus, we can conclude that the "single-norm" Krupka-Weber task might be suitable for situations where only average normative beliefs are the object of interest.

In the third-party punishment task, the modal response across all actions and treatments was not to punish. The most sanctioned action in the DG was "give everything" (give 4 keep 0). In the no charity DG treatment, the average sanctioning is $M = 1.55$ ($SD = 1.78$), and in the DG with charity, we have $M = 1.33$ ($SD = 1.65$). In the allocation game, the most sanctioned actions were the efficient option in the standard version ($M = 1.80$, $SD = 1.76$) and the selfish option in the charity version ($M = 1.77$, $SD = 2.15$).

In the dictator game, participants gave on average 2.7 points ($SD = 0.91$) in the standard version and 2.6 points ($SD = 1.10$) in the charity version. The modal response in both versions was to give 2 points and keep 2 points. The proportion of participants choosing an action that gives more than is kept (i.e., give 3 or 4 points) increases from 1.5% in the standard version to 10.7% in the charity version. In the allocation game, the modal response was the equitable option in the standard version and the maximin option in the charity version. The share of participants choosing the efficient option increases from 3.5% in the standard version to 16.4% in the charity version.

5.2. H1: Non-Randomness of Responses

There are $3^5 = 243$ possible norm drawings for the dictator game scenario, and participants drew 130 unique norms in the standard version and 121 unique norms in the charity version, around one half of all possible combinations. In the allocation game scenario, $3^4 = 81$ drawings were possible, and participants drew 77 unique norms in the standard version and 74 unique norms in the charity version. In all four treatments, the distribution of beliefs was not uniform according to the pre-registered Chi-squared test (all $p < .001$). This result is robust even when excluding those norms that were not drawn by any participant: the

frequency with which participants drew norms was still not uniform, as some norms were drawn overwhelmingly more often than others (all $p < .001$). This finding is also reflected in the number of tokens placed by participants on each norm: the ten norms with the highest average number of tokens account for between 54% (standard dictator game) and 65% (charity allocation game) of the entire distribution of norms in each treatment.

Finding 1: *Consistent with H1, participants' reported beliefs in the norm-drawing task do not follow a random distribution.*

5.3. H2: Multiplicity of Norms

For Hypothesis 2, we look at the frequency of H2A participants who drew a single norm, and H2B participants who drew multiple norms with mutually exclusive appropriate actions. In all treatments, more than half of the participants fall into one of these two categories (Table 2).

Table 2: Proportion of participants drawing a unique norm or multiple norms with mutually exclusive appropriate actions. The 'Others' category includes remaining participants, those who drew two or more norms having at least one appropriate action in common.

Scenario	Charity	Unique	Multiple	Others
DG	Standard	15%	40%	45%
DG	Charity	14%	37%	49%
AG	Standard	17%	37%	46%
AG	Charity	23%	38%	39%

These data confirm that some individuals report only one norm, as it is posed in traditional norm elicitation tasks, but this group represents a minority of the sample. However, it is possible that some participants drew multiple versions of a single norm with different levels of tightness, i.e., all drawings share the same appropriate actions, but the non-appropriate actions vary in their ratings. Even allowing for this possibility, the proportion of participants with a unique norm increases by no more than 2%. Therefore, since the majority of respondents draw multiple norms, it is possible that those who report only one norm are subject to false consensus or pluralistic ignorance. As a second result, we observe that a consistent portion of the sample reports multiple, conflicting norms. This, in turn, may indicate that these participants take conflicting views into account when making their decision.

Finding 2: *Consistent with both H2A and H2B, the majority of participants reported only one norm or multiple contrasting norms, suggesting that there is a multiplicity of views within each scenario.*

5.4. H3: More Punishment with Unique Norm

As the participants predicted by H2A may believe that only a single norm exists, we test whether the probability and rate of punishment of inappropriate actions are higher in these participants than in the rest of the sample. We first focus on the subset of participants who drew a single norm with a single most appropriate action. Both logistic and linear regressions (reported in Appendix C) predicting punishment suggest that these participants are more likely to sanction ($\beta = 0.96$ $[0.05, 1.86]$, $z = 2.072$, $p = 0.038$) and sanction more ($\beta = 0.43$ $[0.05, 0.81]$, $t(822.4) = 2.195$, $p = 0.028$) than other participants. This result is robust to including participants who drew a single norm with multiple appropriate actions (sanction probability: $\beta = 1.07$ $[0.48, 1.66]$, $z = 3.551$, $p < 0.001$; sanction rates: $\beta = 0.41$ $[0.18, 0.65]$, $t(963.4) = 3.419$, $p < 0.001$; Figure 5). If we look at each participant's strongest norm (as measured by the highest number of tokens placed) instead of categorizing participants based on the number of norms drawn, the results remain consistent: the stronger one norm is, the higher the likelihood and magnitude of sanctions (Appendix A.3).

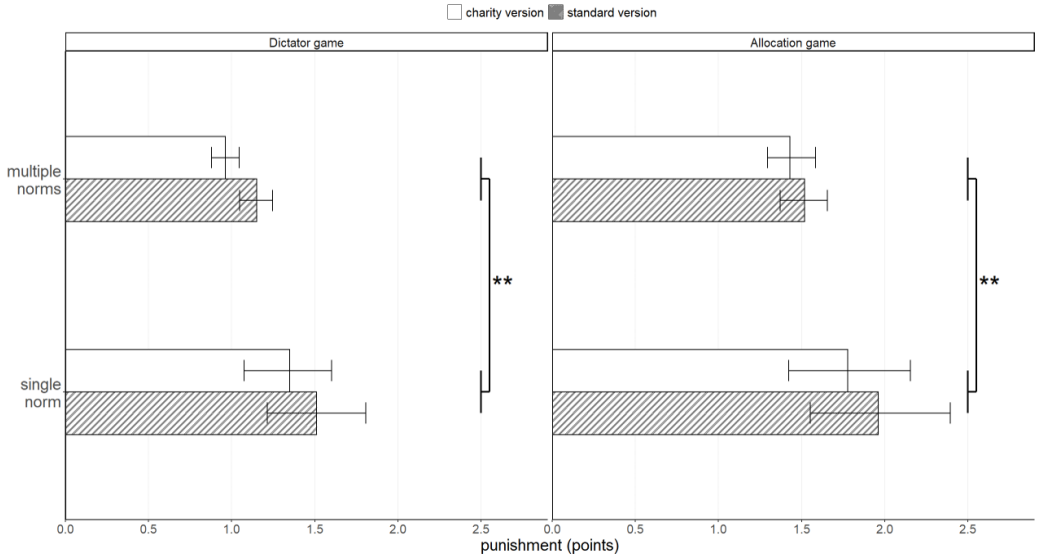

Figure 5: Average punishment, by scenario. Error bars show bootstrapped estimates for the 95% confidence interval around the mean, but are not clustered on the participant level, so should be considered illustrative only. The vertical bars illustrate the significance tests between multiple norm and single norm participants. **: $p < .05$

Finding 3: *Consistent with H3, participants who reported only one norm punished norm violations more frequently and more strongly.*

5.5. H4: Fewer Norms in the Presence of a Charity

This hypothesis predicted that the presence of the charity would have acted as a coordination device and thus would have brought an average reduction in the number of norms drawn.

However, the regression coefficient for charity in the pre-registered Poisson regression turned out to be non-significant ($\beta = 0.00$ $[-0.07, 0.06]$, $z = -0.062$, $p = 0.951$). Indeed, as anticipated in the descriptive statistics, the average number of norms drawn ranges from 4 to 5 in all treatments. This suggests that people do not consider giving to charity as a simpler moral exercise than giving to other human participants.

Finding 4a: *Contrary to H4, we find no significant difference in the number of norms drawn between the standard and charity versions of each scenario.*

Despite this, Figures B.6 to B.9 in Appendix B show that the norms receiving the most tokens differ considerably across treatments. For instance, the standard DG treatment yields norms that favor equality (give 2 keep 2), and the charity DG treatment yields norms that favor generosity (give 4 keep 0). Thus, as an exploratory analysis, we compare the types of norms between the standard and charity versions. We employ a Dirichlet regression for compositional data. To simplify computations, we group together norms that share the same most appropriate actions (i.e., flatter or steeper versions of the same norm). Several norms differ in both scenarios. In dictator game scenarios, the results show that the average number of tokens placed on "give 2 keep 2" norms is smaller in the charity version than in the standard version (6% vs. 31%, $\beta = -0.87$ $[-1.05, -0.68]$, $z = -9.214$, $p < .001$), but conversely the charity version presents a higher share of "give 3 or more" norms (15% vs. 5%, $\beta = 0.44$ $[0.25, 0.62]$, $z = 4.577$, $p < .001$) and "give 4 keep 0" norms (8% vs. 2%, $\beta = 0.33$ $[0.14, 0.52]$, $z = 3.463$, $p = .006$).

These results suggest that people use a different set of norms when dealing with charity as compared to other human beings. It seems that in standard DG, participants mostly consider equality norms, but are unsure about how strong they are. Specifically, Norms 4 and 5 in Figure B.6 have the same most appropriate outcome but differ in how inappropriate the other outcomes are (stricter Norm 5 generates more punishment). In charity DG in Figure B.7, we see only generosity norms (giving everything is the most appropriate outcome). Norms 1, 3, and 4 share the same moral principle, but again are different in terms of strictness (Norm 1 is the strictest and generates the most punishment). Thus, participants agree that generosity is the norm in charity DG but are unsure about how strict the norm is.

In the AG scenario, the charity version presents a smaller share of equitable norms (15% vs 26%, $\beta = -0.34$ $[-0.54, -0.15]$, $z = -3.474$, $p = .008$) or "either equitable or maximin" norms (17% vs 26%, $\beta = -0.27$ $[-0.46, -0.08]$, $z = -2.754$, $p = .047$) compared to the standard version. This suggests again, as in the DG, that charities evoke efficiency norms to a larger degree than other humans.[6]

[6] As a word of caution, we would like to bring the reader's attention to the fact that even though we do find the differences in the types of norms in standard and charity treatments, these types might not be the

Finding 4b: *Participants report different norms in the standard and charity versions of both games. The equality principle seems to dominate norms in the standard treatments. Efficiency norms (in AG) and generosity norms (in DG) dominate in charity treatments.*

6. Discussion

In this study, we presented a new norm elicitation task (the Norm-Drawing Task) that aims to capture the many facets of norm multiplicity: how many norms co-exist in a given environment; how different the beliefs among various people are with regard to this multiplicity; and how multiplicity influences (sanctioning) behavior. We tested the task in two experimental games using two different sets of agents, including or not including a charitable organization.

Results reveal a significant and widespread multiplicity of normative views in all treatments. We also find that a stable proportion of participants, around 15%, believe that only one norm applies in all scenarios, which can be problematic due to possible normative disagreements. Finally—and consistently with our theoretical argument—we find that these "single-norm" participants punish more than their "multiple-norm" counterparts (see discussion below). This suggests that our method is not only able to capture the three above-mentioned desiderata of understanding norm multiplicity but also sheds some new light on the necessity to take this multiplicity into account given that we find on average 4 to 5 different norms perceived by a single participant in all scenarios.

Our results testing hypotheses H1 and H2 suggest that participants do not respond randomly in the task and that most of them indeed perceive a multiplicity of different normative views. The fact that we found multiplicity in very standard games studied for a long time in experimental economics suggests that the possibility for norm multiplicity should be taken into account in all situations, be they experimental or applied. Thus, our new task proposes new avenues for research into this important issue. For example, using our method it is possible to observe instances of false consensus or pluralistic ignorance happening among participants who draw only one norm when instead there is a multiplicity of views in the population.[7] This method also opens up the possibility of discovering intermediate levels of knowledge where participants report multiple norms, but still ignore some or most of them, and their beliefs can be linked to their behavior. In addition to identifying biases in

same in different settings. What we mean is that our participants are aware that their choices in DG and AG can be sanctioned. This changes the behavior as compared to the scenario without punishment. For example, Engelmann and Strobel (2004) find that efficiency and maximin norms are most prevalent in the design without punishment, whereas we find the equality norm as the most frequent. The difference might come from the presence of punishment after the game. See also Appendix A.1.

[7]It is possible to disentangle the two phenomena, but this would require knowledge of personal beliefs since false consensus implies that the norm is one's own, while pluralistic ignorance does not necessarily entail endorsement of such a norm.

recognizing diverse perspectives on a topic, this method also makes it possible to determine whether a norm is perceived as more or less strict by different respondents, adding to the usefulness of this method over traditional approaches.

In addition to enhancing the existing toolbox, our task suggests new possibilities for connecting norm-driven behavior to important societal problems like, for example, violence. Our results suggest that people, who perceive only one norm in an environment where there are many, might get into normative disagreement with others who perceive multiple norms. This is simply because they will not tolerate any form of behavior different from what their solely perceived norm dictates. Thus, the proportion of people who perceive only one norm can serve as an indicator of the potential for normative disagreement or violence.

Our test of hypothesis H3 also reveals how participants who report only one norm tend to punish more and more frequently than other participants, suggesting that the presence of multiple norms may act as a leeway for people in the scenario to act according to their preferences. This finding is in line with the research in experimental and neuro-economics studying related concepts like moral wiggle room (Dana et al., 2007) or moral opportunism (van Baar et al., 2019; Merguei et al., 2022). While it is perhaps not surprising that the perceived presence of multiple norms acts as a deterrent to sanctioning norm violators, it has not been trivial to capture this phenomenon without experimental manipulation (Dimant and Gesche, 2023). Here, we offer a way to study this phenomenon by simply observing the differences in beliefs between respondents.

Our finding that punishment decreases in multi-norm environments (as compared to single-norm environments perceived by some of our participants) also suggests that the multiplicity of norms can have an overall deteriorating effect on norms in general. Less punishment resulting from multiple norms can make norm violations more socially tolerable and thus lead to an erosion of norms. However, it is also possible for multiplicity to increase punishment: although we do not observe it directly in the experiment, multiplicity could lead to the emergence of conflicting norms, such that they are endorsed by different subgroups of the same population. This, in turn, could lead to an increase in sanctions for violations of either norm. These prospects give another reason why it might be important to check for the multiplicity of norms in applied research.

Our analyses concerning the presence of a charity (H4) also provide insight into how perceived norms may change based on certain features of the scenario, such as the agents involved. Although we expected the charity to act as a coordinating device, reducing the number of norms drawn and placing emphasis on the donation towards the charity, our tests for the difference turned out to be non-significant. It is possible that adding a charity might have added an additional layer of complexity to the scenario, rather than removing some perspectives. In support of this possibility, our tests suggest that norms change considerably between the two versions of each scenario.

Based on our experimental findings, we propose that this new elicitation method may be used for several applications. First, because participants can report the existence of multiple beliefs, the method allows researchers and practitioners to study the social landscape in which people believe they are acting. This property is useful in environments that are perceived as polarized: individuals may believe that there are multiple conflicting views and act in consideration of those views. Studying perceived polarization can be helpful in understanding how it affects behavior, for instance in online environments where the perception of extreme views is amplified by the architecture of social media platforms (Bail, 2022). Conversely, given the ability of the method to identify cases of false consensus and pluralistic ignorance, this method may be useful in conjunction with interventions targeting these biases: the method could first determine whether a plurality of beliefs exists and, if so, identify those respondents who report only one or a fraction of the beliefs actually reported by most participants.

Thanks to its flexibility, we can imagine other applications of the norm-drawing task. This method could be used to measure norm "tightness" (Gelfand et al., 2011) in contexts where a single norm is established, such as traffic laws. For example, researchers might include in the task only actions that deviate from the expected rules of behavior (e.g., jaywalking) and ask participants to predict the population's views about these behaviors. In addition to modifying the set of actions studied, researchers may change the number of ratings possible: if for instance researchers are not interested in this tightness-looseness dimension, they could simply ask whether actions are appropriate or inappropriate, and disregard how norm deviations are perceived. This method could be implemented by simplifying the graphical interface, for example by asking participants to mark only the actions they consider appropriate for each view. This approach is particularly useful for reducing complexity in scenarios where there are multiple actions and therefore a combinatorial explosion of possible norms.

References

Artinger, F., Exadaktylos, F., Koppel, H., and Sääksvuori, L. (2010). Applying quadratic scoring rule transparently in multiple choice settings: A note. Working paper.

Aycinena, D., Bogliacino, F., and Kimbrough, E. O. (2022). Measuring norms using the BESA method. Mimeo.

Baader, M. and Vostroknutov, A. (2017). Interaction of reasoning ability and distributional preferences in a social dilemma. *Journal of Economic Behavior & Organization*, 142:79–91.

Bail, C. (2022). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

Barr, A., Lane, T., and Nosenzo, D. (2018). On the social inappropriateness of discrimination. *Journal of Public Economics*, 164:153–164.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, 132:59–72.

Bicchieri, C. and Xiao, E. (2009). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Borel, E. (1921). La théorie du jeu et les équations intégralesa noyau symétrique. *Comptes rendus de l'Académie des Sciences*, 173(1304-1308):58.

Bursztyn, L., Egorov, G., and Fiorin, S. (2020). From extreme to mainstream: The erosion of social norms. *American Economic Review*, 110(11):3522–48.

Centola, D., Becker, J., Brackbill, D., and Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119.

Charness, G., Gneezy, U., and Rasocha, V. (2021). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior & Organization*, 189:234–256.

Cialdini, R. B. and Trost, M. R. (1998). Social influence: Social norms, conformity and compliance.

Crosetto, P. and De Haan, T. (2023). Comparing input interfaces to elicit belief distributions. *Judgment and Decision Making*, 18:e27.

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Dimant, E. (2019). Contagion of pro-and anti-social behavior among peers and the role of social proximity. *Journal of Economic Psychology*, 73:66–88.

Dimant, E. (2023a). Beyond average: A method for measuring the tightness, looseness, and polarization of social norms. *Economics Letters*, 233:111417.

Dimant, E. (2023b). Hate trumps love: The impact of political polarization on social preferences. *Management Science*.

Dimant, E., Clemente, E. G., Pieper, D., Dreber, A., Gelfand, M., and 9, B. S. U. C. H. M. . H. A. . T. P. (2022a). Politicizing mask-wearing: predicting the success of behavioral interventions among republicans and democrats in the us. *Scientific Reports*, 12(1):7575.

Dimant, E., Galeotti, F., and Villeval, M. C. (2023). Motivated information acquisition and social norm formation. Working Paper Available at SSRN: https://dx.doi.org/10.2139/ssrn.4525398.

Dimant, E., Gelfand, M., Hochleitner, A., and Sonderegger, S. (2022b). Strategic behavior with tight, loose, and polarized norms. Working Paper Available at SSRN: https://bit.ly/3ryY3Pc.

Dimant, E. and Gesche, T. (2023). Nudging enforcers: How norm perceptions and motives for lying shape sanctions. *PNAS Nexus*, 2(7):pgad224.

Engelmann, D. and Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American economic review*, 94(4):857–869.

Fragiadakis, D. E., Kovaliukaite, A., and Rojo Arjona, D. (2019). the belief elicitation by superimposition approach. Working paper.

Fromell, H., Nosenzo, D., Owens, T., and Tufano, F. (2021). One size does not fit all: Plurality of social norms and saving behavior in Kenya. *Journal of Economic Behavior & Organization*, 192:73–91.

Gelfand, M., Li, R., Stamkou, E., Pieper, D., Denison, E., Fernandez, J., Choi, V. K., Chatman, J., Jackson, J. C., and Dimant, E. (2021a). Persuading republicans and democrats to comply with mask wearing: An intervention tournament.

Gelfand, M. J., Gavrilets, S., and Nunn, N. (2024). Norm dynamics: Interdisciplinary perspectives on social norm emergence, persistence, and change. *Annual Review of Psychology*, 75.

Gelfand, M. J., Jackson, J. C., Pan, X., Nau, D., Pieper, D., Denison, E., Dagher, M., Van Lange, P. A., Chiu, C.-Y., and Wang, M. (2021b). The relationship between cultural tightness–looseness and covid-19 cases and deaths: a global analysis. *The Lancet Planetary Health*, 5(3):e135–e144.

Gelfand, M. J., Raver, J. L., Nishii, L., Leslie, L. M., Lun, J., Lim, B. C., Duan, L., Almaliach, A., Ang, S., Arnadottir, J., et al. (2011). Differences between tight and loose cultures: A 33-nation study. *Science*, 332(6033):1100–1104.

Groenendyk, E., Kimbrough, E. O., and Pickup, M. (2023). How norms shape the nature of belief systems in mass publics. *American Journal of Political Science*, 67(3):623–638.

Henrich, J. (2017). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the United States. *Annual Review of Political Science*, 22:129–146.

Kimbrough, E. and Vostroknutov, A. (2018). A portable method of eliciting respect for social norms. *Economics Letters*, 168:147–150.

Kimbrough, E. and Vostroknutov, A. (2023a). Resentment and punishment. mimeo, Chapman University and Maastricht University.

Kimbrough, E. and Vostroknutov, A. (2023b). A theory of injunctive norms. mimeo, Chapman University and Maastricht University.

Kimbrough, E. O., Krupka, E. L., Kumar, R., Murray, J., Ramalingam, A., Sánchez-Franco, S., Sarmiento, O. L., Kee, F., and Hunter, R. (2022). On the stability of norms and norm-following propensity: A cross-cultural panel study with adolescents. SSRN Working paper 4025407.

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3):608–638.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

Levy, R. (2021). Social media, news consumption, and polarization: evidence from a field experiment. *American Economic Review*, 111(3):831–70.

Merguei, N., Strobel, M., and Vostroknutov, A. (2022). Moral opportunism as a consequence of decision making under uncertainty. *Journal of Economic Behavior & Organization*, 197:624–642.

Offerman, T., Sonnemans, J., Van de Kuilen, G., and Wakker, P. P. (2009). A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *The Review of Economic Studies*, 76(4):1461–1489.

Peeters, R. and Wolk, L. (2019). Elicitation of expectations using Colonel Blotto. *Experimental Economics*, 22(1):268–288.

Pickup, M., Kimbrough, E. O., and de Rooij, E. A. (2021). Expressive politics as (costly) norm following. *Political behavior*, pages 1–21.

Pickup, M., Kimbrough, E. O., and de Rooij, E. A. (2023). Crossing the line: Evidence for the categorization theory of spatial voting. *British Journal of Political Science*, page 1–11.

Pizziol, V., Demaj, X., Di Paolo, R., and Capraro, V. (2023). Political ideology and generosity around the globe. *Proceedings of the National Academy of Sciences*, 120(15):e2219676120.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roberson, B. (2006). The colonel blotto game. *Economic Theory*, 29(1):1–24.

Sherif, M. (1936). The psychology of social norms.

van Baar, J. M., Chang, L. J., and Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature communications*, 10(1):1483.

Appendix A. Supplementary Analyses

Appendix A.1. Original Hypothesis 4

We report below the original hypothesis 4 as formulated in the pre-registered report, and the related analyses.

Hypothesis: In the standard allocation game, the number of participants guessing a norm with equality rated as the only appropriate action will be smaller than A) the number of participants guessing a norm with the maximin option rated as the only appropriate action, B) the number of participants guessing a norm with efficiency rated as the only appropriate action. To test H4A and H4B we use pairwise post-hoc tests following a Chi-squared test comparing the number of participants drawing each norm type. This test is analogous to the pre-registered test but replaces it because the original model was not identifiable.

Results: In their seminal paper, Dirk Engelmann and Martin Strobel found that equity concerns did not predict choices in most participants. Contrary to this prediction, participants in the standard allocation game treatment demonstrated a strong preference for the equitable option: this action was preferred by 45% of participants. Moreover, the number of participants drawing an equity norm (i.e. a norm with only the equitable option rated as appropriate, $N = 106$) was more frequently drawn than an efficiency norm ($N = 31$) or a maximin norm ($N = 49$, post-hoc Chi-squared test comparisons, all $p < .001$).

Contrary to H4A and H4B, participants drawing an equity norm outnumber those drawing an efficient or maximin norm.

Thus, contrary to expectations, we did not find the same results from the literature suggesting that equity concerns are not the primary normative driver of participants' decisions. Instead, the equity norm was the most frequently drawn compared to the other ones surveyed (maximin, efficiency, selfishness), and by a wide margin. We suggest that the focus on equity may be driven by the particular sample on which we tested our method: prolific participants. Indeed, it is possible that participants in this particular subject pool share a sense of reciprocity to the point that compensation should not be asymmetric, even when participants do not know each other directly. Another factor that may have influenced the observed behavior is the structure of the game compared to the original version: the game in the experiment presents four options that separate the different social preferences taken into account, whereas the original study used a series of games with three options. It is possible that the different payoff matrix and the use of multiple games might have influenced the preferences of players. Running the experiment on different samples may reveal whether norm prevalence depends on the reference group, or on differences between our game and those used in previous studies.

Appendix A.2. Pre-registered test for Hypothesis 3

The original test for H3A and H3B included as independent variables the participant's minimum appropriateness rating for the given action across all guessed norms (inappropriate, neither, appropriate), and the interaction between this variable and the presence of a single norm with a unique appropriate action. This test split results between "inappropriate" and "neither appropriate nor inappropriate" ratings, whereas the hypothesis does not distinguish between the two cases. We report for completeness the results of the tests here, which are in line with the tests reported in the main text. Results confirm the finding that reporting a unique norm with a unique appropriate action in the norm-drawing task yields higher likelihood to sanction (H3A: $\beta = 1.15$ $[0.23, 2.07]$, $z = 2.449$, $p = 0.014$) and to greater sanctions (H3B: $\beta = 0.51$ $[0.12, 0.89]$, $z = 2.551$, $p = 0.011$). As an exploratory test, we also examine the effect of our main independent variable divided by the rating. This analysis suggests that the increased likelihood and amount sanctioned is mainly driven by actions rated as "neither appropriate nor inappropriate" (H3A: $\beta = 1.93$ $[0.83, 3.02]$, $z = 3.452$, $p_{fdr} = 0.001$; H3B: $\beta = 0.63$ $[0.17, 1.09]$, $z = 2.680$, $p_{fdr} = 0.015$). The contrasts for "inappropriate ratings is in the same direction but does not reach significance (H3A: $\beta = 0.37$ $[-0.65, 1.38]$, $z = 0.707$, $p_{fdr} = 0.480$; H3B: $\beta = 0.38$ $[0.17, 1.09]$, $z = 1.686$, $p_{fdr} = 0.092$). These results suggest that participants who consider only one action to be appropriate are less tolerant of gray areas, such as actions that are neither generally inappropriate nor appropriate.

Appendix A.3. Hypothesis 3 using the number of tokens

As an additional robustness test for H3, we tested whether the rate and magnitude of sanctions were proportional to the number of tokens placed on the norm that received the most tokens. To do so, we ran the same regressions as in H3. As in the original regressions, we included only actions that were rated as "inappropriate" or "neither appropriate nor inappropriate," in this case according to the norm that received the most tokens. Instead of the original dummy variable, we included the number of tokens placed on that norm. In the case of ties between multiple norms, participants were excluded from this exploratory analysis. The results again are in line with the ones reported in the main text: the higher the number of tokens placed, the higher the likelihood of sanctioning (H3A: $\beta = .010$ $[.002, .019]$, $z = 2.899$, $p = 0.017$) and the amount sanctioned (H3B: $\beta = .005$ $[.001, .009]$, $t(593.7) = 2.899$, $p = 0.004$). According to this last model, for every 10 more tokens placed on the norm with the most tokens, sanctions increase by an average of 0.05 points.

Appendix B. Additional Figures

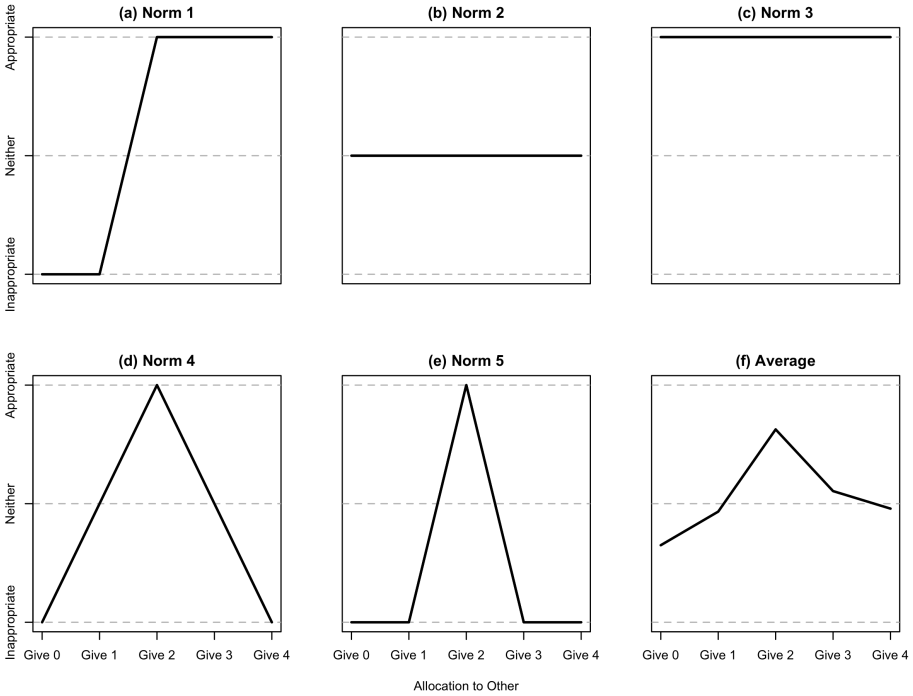

Figure B.6: Top 5 Most Commonly Drawn Norms, and the Belief-Weighted Average Norm, Standard DG.

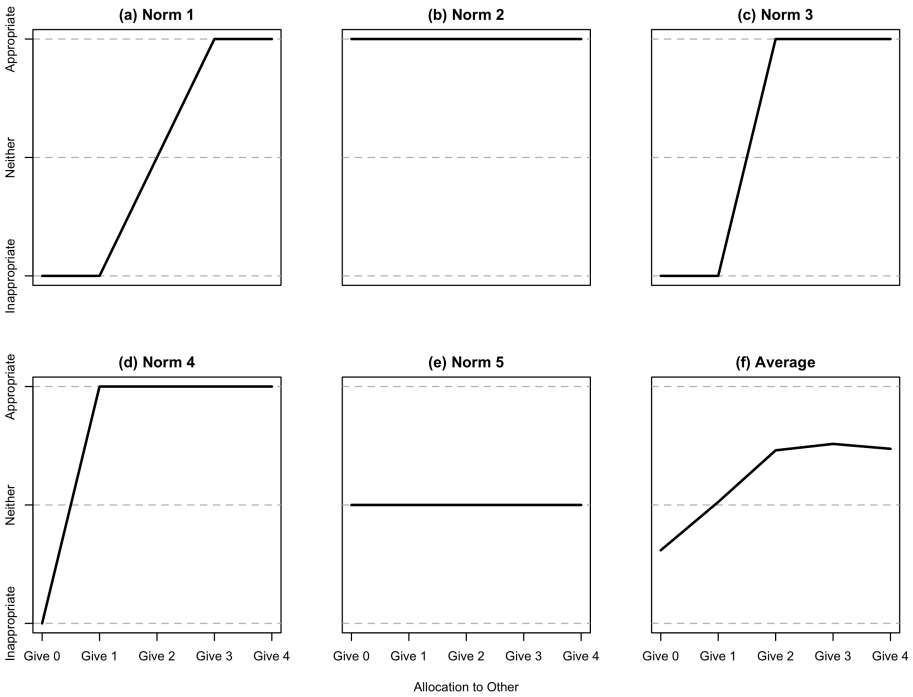

Figure B.7: Top 5 Most Commonly Drawn Norms, and the Belief-Weighted Average Norm, Charity DG.

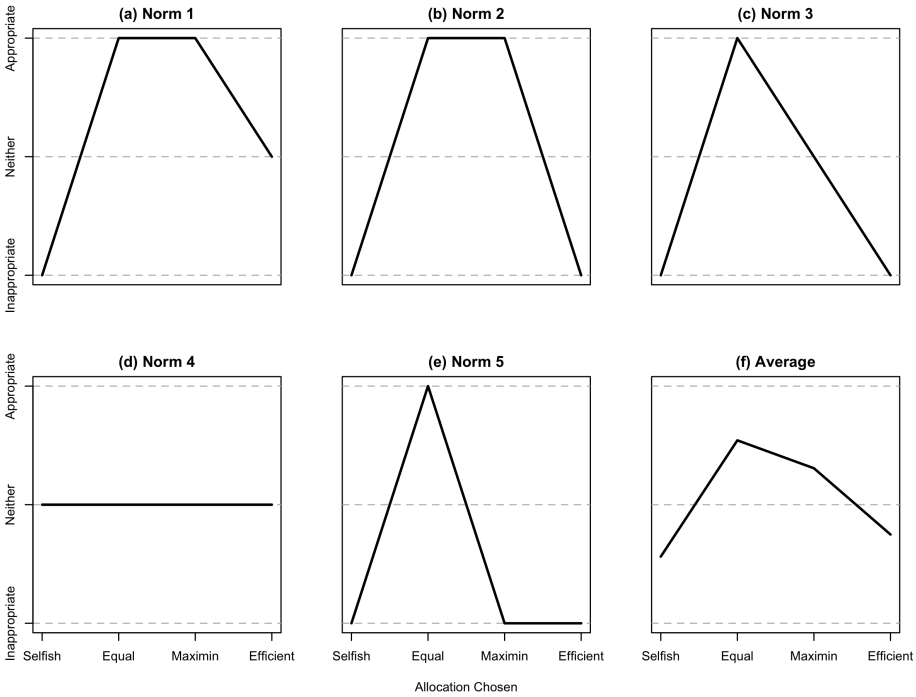

Figure B.8: Top 5 Most Commonly Drawn Norms, and the Belief-Weighted Average Norm, Standard AG.

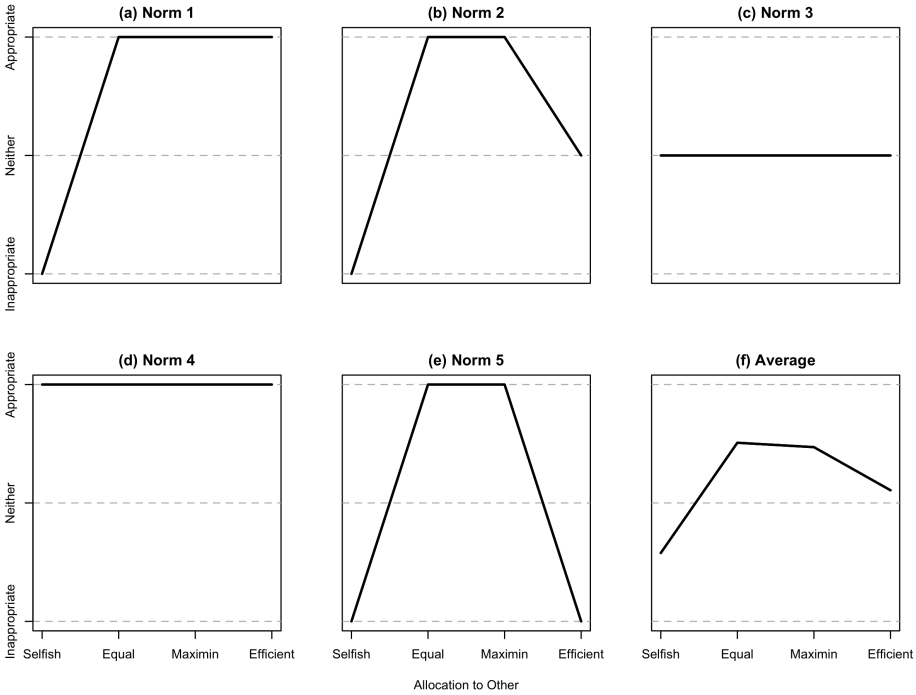

Figure B.9: Top 5 Most Commonly Drawn Norms, and the Belief-Weighted Average Norm, Charity AG.

Appendix C. Supplementary Tables

	punishment > 0		
Predictors	*β*	*95% CI*	*p*
(Intercept)	0.90	0.64 – 1.26	0.552
Single appropriate action	2.61	1.05 – 6.47	**0.038**
Charity	0.76	0.52 – 1.12	0.162
Allocation game	1.35	0.92 – 1.99	0.129
Random Effects			
$N_{independent\ observations}$	811		
$N_{observations}$	3188		
Marginal R^2 / Conditional R^2	0.009 / 0.635		

Supplementary Table 1. Results summary for the logistic mixed-effects regression testing H3A. Intercept: standard Dictator game, participant who drew at least two norms or a norm with multiple appropriate actions; Single appropriate action: participant who drew a single norm with a single appropriate action; Charity: effect of charity; Allocation game: effect of the allocation game scenario. Coefficients are on the odds ratio scale.

	punishment		
Predictors	*β*	*95% CI*	*p*
(Intercept)	1.15	1.01 – 1.29	**<0.001**
Single appropriate action	0.43	0.05 – 0.81	**0.028**
Charity	-0.13	-0.30 – 0.03	0.113
Allocation game	0.45	0.28 – 0.61	**<0.001**
Random Effects			
$N_{independent\ observations}$	811		
$N_{observations}$	3188		
Marginal R^2 / Conditional R^2	0.025 / 0.423		

Supplementary Table 2. Results summary for the linear mixed-effects regression testing H3B. Intercept: standard Dictator game, participant who drew at least two norms or a norm with multiple appropriate actions; Single appropriate action: participant who drew a single norm with a single appropriate action; Charity: effect of charity; Allocation game: effect of the allocation game scenario.

	punishment > 0		
Predictors	*β*	*95% CI*	*p*
(Intercept)	0.83	0.59 – 1.19	0.316
Single norm	2.92	1.62 – 5.27	**<0.001**
Charity	0.73	0.49 – 1.10	0.133
Allocation game	1.32	0.88 – 1.98	0.179
Random Effects			
$N_{independent\ observations}$	811		
$N_{observations}$	3188		
Marginal R^2 / Conditional R^2	0.017 / 0.640		

Supplementary Table 3. Results summary for the logistic mixed-effects regression testing H3A, single norm and any number of appropriate actions. Intercept: standard Dictator game, participant who drew at least two norms; Single norm: participant who drew a single norm; Charity: effect of charity; Allocation game: effect of the allocation game scenario. Coefficients are on the odds ratio scale.

	punishment		
Predictors	*β*	*95% CI*	*p*
(Intercept)	1.12	0.98 – 1.27	**<0.001**
Single norm	0.41	0.18 – 0.65	**0.001**
Charity	-0.14	-0.31 – 0.02	0.088
Allocation game	0.43	0.27 – 0.60	**<0.001**
Random Effects			
$N_{independent\ observations}$	811		
$N_{observations}$	3188		
Marginal R^2 / Conditional R^2	0.028 / 0.423		

Supplementary Table 4. Results summary for the linear mixed-effects regression testing H3B, single norm and any number of appropriate actions. Intercept: standard Dictator game, participant who drew at least two norms; Single norm: participant who drew a single norm; Charity: effect of charity; Allocation game: effect of the allocation game scenario.