TI 2016-023/I
Tinbergen Institute Discussion Paper

Everybody's doing it: On the Emergence and Persistence of Bad Social Norms

David Smerdon[1,2]
Theo Offerman[1,2]
Uri Gneezy[3]

[1] *Faculty of Economics and Business, University of Amsterdam, the Netherlands;*
[2] *Tinbergen Institute, the Netherlands;*
[3] *UC San Diego, United States.*

Tinbergen Institute Amsterdam
Gustav Mahlerplein 117
1082 MS Amsterdam
The Netherlands
Tel.: +31(0)20 525 1600

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

"Everybody's Doing It": On the Emergence and Persistence of Bad Social Norms

David Smerdon[1,2], Theo Offerman[1,2] and Uri Gneezy[3]

[1]University of Amsterdam
[2]Tinbergen Institute, Amsterdam
[3]UC San Diego

April 2016

Abstract

Social norms permeate society across a wide range of issues and are important to understanding how societies function. In this paper we concentrate on 'bad' social norms - those that are inefficient or even damaging to a group. This paper explains how bad social norms evolve and persist; our theory proposes a testable model of bad norms based on anecdotal evidence from real-world examples. We then experimentally test the model and find empirical support to its main predictions. Central to the model is the role of a person's *social identity* in encouraging compliance to a norm. The strength of this identity is found to have a positive effect on bad norm persistence. Additionally, while the size of the social group does not have a long run effect, smaller groups are more likely to break a bad norm in the short term. Furthermore, the results suggest that both anonymous communication and increasing information about others' payoffs are promising intervention policies to counter bad norms.

1 Introduction

Social norms are central to our understanding of behavior. They provide informal rules that govern our actions within different groups and societies and across all manner of situations, from a simple handshake or queuing in a line to taking revenge or engaging in courtship. Powerful and pervasive, the gravity of the effects of social norms can range from the slight (such as fashion or restaurant etiquette) to the dire (such as committing female genital mutilation or murder). Social norms are discussed in all social sciences, and economic modeling has made a useful contribution to the extant discussion in recent decades. Rational choice models and tools from game theory have helped to frame positive, welfare-enhancing social norms as effective and pragmatic means of solving coordination problems with multiple equilibria (Young, 2008). Such norms develop in order to overcome market failure, mitigate negative externalities or promote positive ones so as to facilitate some collective goal (Arrow, 1970). Evolutionary economists model this development in terms of 'evolutionary stable strategies', concluding that only socially beneficial (or 'good') norms are likely to emerge (Hechter & Opp, 2001).

However, social norms that are inefficient from a welfare perspective do exist in the real world and are worthy of serious attention. The destructive potential of social norms has been of interest to academics in various disciplines for some time, particularly among the functionalism schools of psychology and sociology. One need only recall the famous Milgram (1974) obedience experiments, the Stanford prison experiments (Zimbardo, 1972) or the Asch (1956) conformity experiments to appreciate the power that social pressures can have on individual rationality. Historical norms, such as the custom of dueling in the American South (Lessig, 1995) and a millennium of female foot-binding in China (Mackie, 1996), have shown extremely high levels of persistence. In modern times, economic literature has highlighted the important role that bad social norms play in many topical policy issues, such as in environmental policy (Kinzig *et al.* , 2013), human rights reform (Prentice, 2012), and many issues in development economics, such as income inequality (Singh & Dhumale, 2000), population growth (Munshi & Myaux, 2006) and HIV/AIDS (Young *et al.* , 2010).

Cognitive theories from psychology and sociology and rational choice models of equilibrium selection have struggled to provide a cohesive explanation of how such social norms that damage welfare can possess such stubbornness and longevity, despite the apparently obvious disadvantages to society and its individuals. As Eggertsson (2001, p. 78) writes, "Economists have good reason to reconsider their theories and methods if they are unable to explain the existence and persistence of inefficient norms." A better understanding of this underexplored topic would therefore offer not only a theoretical contribution but have the potential to impact a wide range of practical applications that concern the welfare of individuals and groups.

In this paper, we conjecture that bad norms initially emerge as good norms, but changing con-

ditions over time alter the payoff structure such that the norm not only ceases to solve negative externalities, but actually begins to promote them. The historical salience of the norm results in the corresponding behavior being hardwired into individuals' social identity such that it becomes problematic for the group to coordinate on some other, norm-inconsistent choice. The stronger the sense of identity in relation to the behavior, the more likely it is that the bad norm can persist.[1]

The primary objective of this paper is to demonstrate the conditions under which bad social norms can emerge and persist within a group. We argue that the two key features required for the evolution and persistence of inefficient social norms are a shift in incentives over time and a strong sense of group identity. In this, our paper is closely related to the literature on social interactions. Within this literature we draw chiefly on the theoretical approach of Brock and Durlauf (2001), which can be considered one of the gold standards for modeling the behavior of groups with social interaction effects. Their model of discrete choice considers noncooperative agents whose actions are interconnected with the payoffs of other group members. In dynamic environments in which each individual is faced with a binary choice that will affect others through collective social utility functions, they show that multiple locally stable equilibrium levels of average group choice can exist, dependent on the (relative) strength of social interactions on utility. Aggregate group behavior in their model stabilizes around a common choice; the welfare-maximizing choice is always a locally stable equilibrium, while local stability of the welfare-inefficient choice requires large social utility effects. We extend this theoretical approach to study social norms, framing interaction effects in the context of social identity.

Our paper makes several contributions to the literature on social norms and on social interactions more broadly. First, our theoretical approach differs from Brock and Durlauf (2001) in two key respects. Individuals in our model are uncertain about the true expected private payoffs to other members of the group, which allows for heterogeneity of expectations. This extension permits an application to contexts in which this heterogeneity leads to *pluralistic ignorance*. Pluralistic ignorance refers to a situation in which most individuals in a group have a positive personal incentive to deviate from the norm, but believe that the majority of group members have a private incentive to keep to the status quo. Our model allows us to simulate environments exhibiting this effect, which has been linked to the propagation of various damaging social issues, such as college binge-

[1] By way of a practical example, consider handshaking. Shaking hands as a form of greeting is believed to have originated around 2,000 years ago between opposing military personnel (D'Cruz, 2005). It served as a signaling mechanism that the offeror was not concealing a weapon. Particularly during wartime in medieval societies, the small personal effort of the physical act was easily outweighed by the mutual benefits of ensuring peaceful discourse. The custom spread and today has become a very strong social norm in Western culture, although sending a signal that an individual is unarmed no longer carries the same importance. However, hand-to-hand contact is also recognized as one of the main channels for common infections; the H1N1 epidemic of 2009 led many school administrators in the United States to ban handshaking at graduation ceremonies in that year, and more recent influenza scares prompted the 2012 British Olympic team to shun this standard act of sportsmanship before events (Neyfakh, 2013). Yet, despite these isolated instances of imposed non-conformity and the efforts of small activist groups such as the website www.StopHandshaking.com, the norm remains a bastion of modern etiquette.

drinking (Schroeder & Prentice, 1998), tax avoidance (Wenzel, 2005), school bullying (Sandstrom *et al.* , 2013) and the spread of HIV/AIDS due to stigmas against condom usage (Gage, 1998). A corollary of this approach is that we describe a dynamic behavioral rule that provides insight into expectations formation as well as suggests which equilibrium is selected, and when. A second, more technical, extension is that we generalise the results of Brock and Durlauf, which are targeted at econometric implementation, to all symmetric shock distributions.

Secondly, our experimental results shed light on our and other theoretical attempts to model these phenomena. Our first main empirical result is that the stronger the social identity of a group, the more likely a bad norm is to persist. This result is perhaps not surprising, but it is important because it highlights a necessary requirement for the emergence and persistence of bad norms. Our second empirical result relates to group size and is subtler in nature. We find that smaller groups are better at breaking bad norms in the short term, but across longer horizons, these effects disappear given the same relative strength of social identity.

Having established the fundamental conditions for bad norm to exist, our final experimental results reflect attempts to break down their persistence. We find experimental support for two promising interventions: increasing information about common utility and introducing communication. The success of these treatments against bad norms is surprising and could not be predicted *ex ante* from the model, and suggest implications for social policymakers.

A key feature of our paper is the assimilation of social identity theory with existing models of social interactions, incorporating the relative importance to an individual of group conformity into the utility function. The idea of one's sense of self, or identity, affecting behavior is not new; the concept of social identity has been known to psychologists since it was pioneered in the 1970s by Henri Tajfel and John Turner. The main assumption of this theory is that group membership acts both to build up a sense of identity and to bolster self-esteem, and thus individuals favor behavior that reaffirms the self-concept (Tajfel & Turner, 1986, 1979). Akerlof and Kranton (2000) were the first economists to attempt to explicitly model this concept by incorporating identity into an individual's utility function. Through their theory they show that the outcome of various problems both with and without social interactions can be quite different from that predicted by standard economic models. A raft of recent empirical evidence has since demonstrated that social identity can influence individual decision-making and behavior in a wide range of respects, such as polarization of beliefs (Hart & Nisbet, 2011; Luhan *et al.* , 2009), preferences over outcomes (Charness *et al.* , 2007), trust (Hargreaves Heap & Zizzo, 2009), redistribution preferences (Yan Chen, 2009), punishment behavior (Abbink *et al.* , 2010), discrimination (Fershtman & Gneezy, 2001), self-control (Inzlicht & Kang, 2010), competitiveness (Gneezy *et al.* , 2009) and time horizons for decision-making (Mannix & Loewenstein, 1994).

An important assumption of social identity theory, which we too adopt, is that people can derive

identity-based utility both from their own actions and from others' actions, so long as these actions support their sense of self. In this paper, we focus on situations in which identity is cultivated through a majority of a group coordinating on a set action. A bad norm is then defined as one in which a plurality would prefer the group to shift to a socially (and largely individually) preferable behavior, but coordination issues prevent such a switch.[2]

The formulation of a testable model of social norms that can accommodate the evolution and persistence of inefficient group behavior and whose predictions are confirmed both in simulations and in the laboratory is important. This accord of theory and experimental results has proven difficult for economic models that embody social interactions in which "there are few, if any, restrictions on equilibrium behavior and, hence, such models have little or no predictive power" (Postlewaite, 2010, p. 33).

The remainder of the paper is organized as follows. Section 2 presents the theory. It shows how identity and social interactions shape a unified, tractable theory of bad norms. The model's implications are derived both analytically and through simulation. In Section 3, we detail the design and procedure used to transpose the model into the laboratory. Section 4 then discusses the experimental results, from which the conditions under which bad norms can evolve and persist are demonstrated. Finally, Section 5 discusses the tractability of the findings and proposes research streams for extension.

2 Theory

2.1 The Game

We investigate the norm formation process in the 'Identity Game'. In this game, N players repeatedly choose between two options over a number of rounds. An individual's payoff from the chosen option in each round is composed of utility from both her *private value* and her *social value*, which measures the congruence between the individual's choice and those of the group. In each round, every player is informed of her private values of the two options. There is uncertainty about the private values pertaining to the other players, but each player knows that everyone's values are positively correlated. Specifically, it is known that, for each round, each option's private value is

[2] A related literature shows how bad norms can emerge in team production processes that are characterized by a minimum effort production function. In the minimum effort game, players simultaneously exert costly effort, and the minimum effort in the team determines its productivity. The stage game hosts a multitude of Pareto ranked equilibria. In experiments, subjects usually quickly coordinate on a bad equilibrium that offers them a secure but low payoff, unless group size is very small (Van Huyck *et al.* , 1990; Knez & Camerer, 1994). It appears to be surprisingly hard to avoid bad outcomes in minimum effort games, but there are some reliable factors that help subjects coordinate on better outcomes (Cachon & Camerer, 1996; Weber, 2006; Chaudhuri *et al.* , 2009; Kopanyi-Peuker *et al.* , 2015). Our interest is in studying bad norms in applications beyond the labor market. The contribution of our paper is that we show how bad norms can arise, persist and be broken in a completely different class of games.

comprised of the sum of a *common value* and an individual-specific *private shock*, for which the (continuous) distribution is known. It is also known that new private shocks are drawn every round, and that the (unobserved) common values can change across rounds. Therefore, it may be that the initially 'good' option (i.e. the option possessing the higher common value) loses its attractiveness and becomes the 'bad' option after some time. In line with the approach of Brock and Durlauf (2001), an individual i receives in a given round a payoff of:

$$V(\omega_i) = u(\omega_i) + S(\omega_i, \omega_{\neg i}) + \epsilon_i(\omega_i), \qquad \omega_i \in \{-1, 1\} \tag{1}$$

Here, ω represents the choice variable, taking the value of -1 or 1. $u(\omega_i)$ represents the common value from i's choice ω_i, and $\epsilon_i(\omega_i)$ is an individual choice-dependent shock. Individuals in this game do not separately observe the common values or their individual shocks, but rather the combined private value $v_i(\omega_i) = u(\omega_i) + \epsilon_i(\omega_i)$.

The individual shocks $\epsilon_i(\omega_i)$ are identically and independently distributed across all individuals and choices such that the difference $\epsilon_i(-1) - \epsilon_i(1)$ has a known probability distribution function $F(\cdot)$. Its density is symmetric around a mean of zero.

$S(\omega_i, \omega_{\neg i})$ gives the social value of the choice. In this game, the assumption is made that the utility derived from group identity exhibits "constant and totalistic strategic complementarity" (Brock & Durlauf, 2001, p. 238). This means that individuals are always happier by the same amount when one more person makes the same choice as them. With this assumption, the form of social value is stipulated in (2):

$$S(\omega_i, \omega_{\neg i}) = J\omega_i m_i \tag{2}$$

where $m_i = \frac{\sum_{j \neq i} \omega_j}{N-1}$ represents the average choice of the *other* subjects, and $J(> 0)$ represents the *identity factor*, which weights social utility relative to the direct private-value payoff[3].

2.2 Equilibria of the Identity Game

We are interested in the expected average choice of the group, $m^* = \frac{\sum_{i=1}^{N} \omega_i}{N}$. In the remainder, we define an equilibrium ρ^* of the Identity Game as *the expected proportion of the group choosing*

[3]Brock and Durlauf (2001) also discuss a second social utility function in their paper, of the form $S(\omega_i, \omega_{\neg i}) = \frac{J}{2}(\omega_i - m_i)^2$. The equilibrium analysis that follows is identical in this case. As the authors themselves show, the second form can be rewritten as $J\omega_i m_i - \frac{J}{2}(1 + m_i^2)$ in order to show that the portion of social utility containing the choice variable ω_i is the same for both functional forms. Therefore, an individual maximizing expected utility follows the identical rule 'Choose $\omega_i = -1$ if $d_i > 2Jm_i^e$' in both cases.

$\omega_i = -1$, *such that no individual would be better off changing her choice in expectation.* The equilibrium is therefore specified by:

$$\rho^* = \frac{1 - m^*}{2} \tag{3}$$

Individuals cannot *ex ante* observe m_i but instead must base their decision on an expectation of average group choice:

$$m_i^e = \frac{\sum_{j \neq i} \mathbb{E}_i(\omega_j)}{N - 1}$$

where $\mathbb{E}_i(\omega_j)$ represents i's expectation over j's choice. In equilibrium, individuals' expectations are consistent with how others play the game. It is convenient to define $d = u(-1) - u(1)$ as the difference in common values and $d_i = v_i(-1) - v_i(1)$ as the difference in private values for individual i. We are interested in situations in which social utility is expected to play a role for the majority of individuals; that is, $-2J \leq d \leq 2J$. Moreover, for the stage game it is assumed that individuals know both the distribution generating the private shocks for all individuals and the common values for each choice[4].

Proposition 1. *If all individuals follow a common threshold decision rule "Choose $\omega_i = -1$ if $d_i > c^*$" for some common threshold c^*, and private shocks follow a known, symmetric distribution, then an equilibrium expected average choice level of the group, m^*, solves:*

$$m^* = 2F(2Jm^* - d) - 1$$

where F is the CDF of the difference in private shocks.

It follows from the decision rule specified in 1 that, in equilibrium, we require that players prefer $\omega_i = 1$ at least as much as $\omega_i = -1$ if $d_i < c^*$, that players prefer $\omega_i = -1$ at least as much as $\omega_i = 1$ if $d_i > c^*$ and, in particular, that a player is exactly indifferent between $\omega_i = -1$ and 1 if she draws private values with a difference equal to the threshold c^*. We use this latter property of the equilibrium to endogenously calculate the threshold.

The threshold c^* depends both on an individual's beliefs about group behavior as well as the (fixed) identity strength. Solving for this threshold allows us to compute a general equilibria condition that holds for any symmetric distribution of the private shocks[5]. Then an individual i

[4]For example, individuals may have come to know the common values from historical information or experience.

[5]Here we expand upon Brock and Durlauf (2001), who assume that shocks follow an extreme value distribution. The authors make use of some convenient properties of this distribution to analytically compute rational expectations equilibria from the symmetry of N expectations equations.

maximizing her expected utility chooses $\omega_i = -1$ if $d_i > 2Jm_i^e$. To endogenously solve for an equilibrium, we first rewrite m_i^e as:

$$m_i^e = \frac{1}{N-1}\sum_{k=0}^{N-1}\left(\binom{N-1}{k}p^k(1-p)^{(N-1-k)}(2k-N+1)\right) \tag{4}$$

where p is the probability of a single draw of $d_i < c^*$ so that i chooses $\omega_i = 1$. Then each term in the series is the expected value for each possible value of m_i, which can be written in the form $\frac{2k-N+1}{N-1}$ for each $k \in \{0, N-1\}$.

Letting m_i^{e*} be the equilibrium expected average choice of the others in a group, corresponding to a threshold c^*, we can rewrite $c^* = 2Jm_i^{e*}$ in (4). Then solving for an individual i drawing exactly $d_i = c^*$ with $V(-1) = V(1)$ allows us to solve endogenously for the expectation $m_i^{e*} = m_j^{e*} \ \ \forall i, j$:

$$m_i^{e*} = \frac{1}{N-1}\sum_{k=0}^{N-1}\binom{N-1}{k}F(2Jm_i^{e*}-d)^k(1-F(2Jm_i^{e*}-d))^{(N-1-k)}(2k-N+1) \tag{5}$$

At first sight, an individual's expectations appears to depend on the size of the group, N. We perform the replacements $M = N-1$ and $F = F(2Jm_i^{e*}-d)$ for notational convenience to rewrite (5) as:

$$\frac{1}{M}\sum_{k=0}^{M}\binom{M}{k}F^k(1-F)^{(M-k)}(2k-M)$$

It can be shown that the sum $\mathcal{S}_M$ of this series is independent of group size as follows:

$$\begin{aligned}
\mathcal{S}_M =& \frac{1}{M}\sum_{k=0}^{M}\binom{M}{k}F^k(1-F)^{(M-k)}(2k-M)\\
=& \frac{2}{M}\sum_{k=0}^{M}\left(k\binom{M}{k}F^k(1-F)^{(M-k)}\right)-1\\
=& \frac{2FM}{M}\sum_{j=0}^{M-1}\left(\binom{M-1}{j}F^j(1-F)^{(M-1-j)}\right)-1 \qquad \text{(letting } j=k-1\text{)}\\
=& 2F-1
\end{aligned} \tag{6}$$

Thus, (5) can be rewritten as $m_i^{e*} = 2F(2Jm_i^{e*}-d)-1$. The researcher's prediction of the expected average choice level of the whole group is then (5) with the limits extended from $N-1$ to N. As we have seen, the sum of the series is not dependent on N and so the expected average

choice level of the group, m^*, solves:

$$m^* = 2F(2Jm^* - d) - 1 \tag{7}$$

(7) is therefore the stage-game equilibria condition for the expected average choice level, corresponding to a common threshold c^*, for *any* symmetric, continuous distribution of shocks, generalizing Brock and Durlauf (2001). Consider the case where $\epsilon_i(\omega_i) \sim \mathcal{N}(0,1)$, so that the difference $\epsilon_i(-1) - \epsilon_i(1) \sim \mathcal{N}(0,2)$. Then the CDF of this distribution is $F(X) = \Phi(\frac{X}{\sqrt{2}})$. We can rewrite this in terms of the error function by using $\Phi(X) = \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}(\frac{X}{\sqrt{2}})$, so that (7) becomes:

$$m^* = \mathrm{erf}(\frac{2Jm^* - d}{2}) \tag{8}$$

(8) cannot be solved analytically, but a numerical analysis reveals the existence of either one or three equilibria for different values of d and J, which coincides with Brock and Durlauf's (2001) analysis of the extreme value distribution. Multiple equilibria exist only when J is sufficiently large relative to d. In such cases, and adopting for convenience the notation of (3), two stable equilibria close to the poles $\rho^*_- \approx 0$ and $\rho^*_+ \approx 1$ emerge[6]. With some abuse of terminology, an unstable 'mixed-proportions' equilibrium $\rho^* \in (\rho^*_-, \rho^*_+)$ also exists. When the identity factor is not sufficiently large with respect to the difference in common values, there exists a unique equilibrium close to the 'good' pole (i.e. $\rho^* \approx 1$ when $d > 0$).

2.3 Dynamic Analysis

The analysis of the stage game corroborates Postlewaite's (2010) criticism of models of social interactions in that multiple equilibria can exist with no guidance as to how to predict their likelihood. In a repeated setting of cases in which three equilibria exist in our model, the mixed-proportioned equilibrium not only disappears in expectation, but is also unstable in the unlikely event of its realization; new draws of private shocks in subsequent rounds trigger a 'snowball effect' whereby individual decisions quickly converge towards local stability near one of the poles. We now describe a dynamic process that allows a researcher to predict the likelihood of the emergence of these two stable equilibria under different conditions.

Our dynamic model differs from that of Brock and Durlauf (2001) in the sense that here individ-

[6]Recall that ρ^* is the expected proportion of the group choosing $\omega_i = -1$. Due to the continuous distribution of the private shocks across all possible values on the real axis, there is always a positive probability of a private difference $|d_{it}| > 2J$, and so the equilibrium proportions are never exactly at the poles 0 and 1. It is noteworthy that it is not required that all or even any of the individuals have a private value preference for a particular choice for it to exist as a pure equilibrium.

uals are not informed of the common values and therefore the distribution generating the private values. Equilibrium arises when the common values are stable for some time, and individuals can learn to forecast how others actually behave; however, we must specify some assumptions about individuals' belief formation across rounds. Time subscripts are now introduced into the notation in order to describe the dynamic environment. We assume that in a given round t, players form their expectations about the rest of the group's behavior, m^e_{it}, via a common function ψ, which depends on the only two pieces of information available to individuals: the difference in their private values, and a common group 'norm'. When $|d_{it}|$ exceeds $2J$, individual i's private value difference is so high that she no longer considers social interactions at all, and so restrictions on expectations for our purposes need only address ψ for the range $d_{it} \subset [-2J, 2J]$.

A plausible and parsimonious function for the formation of individuals' expectations in round t is:

$$\psi(d_{it}, m_t) = \delta m_{t-1} - (1-\delta)\frac{d_{it}}{2J}, \qquad \delta \in [0,1] \tag{9}$$

Here, m_{t-1}, the group choice of the previous period, represents a simplified form of a common norm, and δ represents how an individual weighs the new information stemming from her private values against this group norm. Under this specification, an individual's expectations of the proportion of the group choosing $\omega = -1$ in a given round are decreasing in her private value difference d_{it} and positively skewed in the direction of the common norm. In contrast to Brock and Durlauf's (2001) setup, individuals will therefore have different expectations about the behavior of others, depending on the realisation of their own private values.

Consider a period of rounds in which the difference in the common values, d_t, is relatively stable. $\psi(d_{it}, m_t)$ can be thought of as a belief-updating process that guides individuals' choices towards a stable, long-run 'equilibrium proportion' choosing $\omega_{it} = -1$. When $|d_t|$ is very small (relative to J), the system moves faster towards equilibrium for high δ because individuals are congregated by the existing norm, although the equilibrium may not be the socially optimal choice. The current norm helps individuals overcome their coordination difficulties, but in doing so can entice the group to forego potential social welfare. When $|d_t|$ is very large, the system can stabilize quickly even for low δ, as normative effects are not needed for coordination on the superior choice.

The expectation formation process (9) enables a researcher who knows the common values and the distribution of the private shocks (though not their realizations) to predict both the average group choice m_t in a given round and, if the common values remain constant, the dynamically-stable equilibria over a period.

Proposition 2. *If individuals form expectations of group behavior according to (9) and the difference in common values is constant for a sufficiently long period of time so that one can write $d_t = d$,*

then a stable equilibrium expected average group choice at the end of the period solves:

$$m^* = 2F\left(\frac{2J\delta}{2-\delta}m^* - d\right) - 1 \tag{10}$$

The proof is similar to the equilibrium condition analysis in the stage game leading to (7). Recall that an individual i does not know the common values and thus the distribution of private values from which those of the other group members are drawn. Substituting (9) into the threshold decision rule, i chooses $\omega_{it} = -1$ if $d_{it} > 2J\left(\delta m_{t-1} - (1-\delta)\frac{d_{it}}{2J}\right)$, which can be rewritten as:

$$d_{it} > \frac{2J\delta}{2-\delta}m_{t-1}$$

Following similar sum-of-series calculations to (5) leads to an equilibrium average group choice prediction in a given round t of the form:

$$m_t^* = 2F\left(\frac{2J\delta}{2-\delta}m_{t-1} - d_t\right) - 1$$

For long-run stability to occur during a period in which $d_t = d_{t+1} = d$ we replace $m_{t-1} = m_t = m^*$ in expectation, which leads immediately to (10).

We now turn to the question of whether a bad norm can persist in a dynamic setting. First, we rewrite (10) in terms of the equilibrium proportion of the group choosing $\omega_i = -1$ at the end of the period:

$$\rho^* = F\left(d - \frac{2J\delta(1-2\rho^*)}{2-\delta}\right) \tag{11}$$

Let the private shocks again be normally distributed according to $\epsilon(\omega_{it}) \sim \mathcal{N}(0,1)$. Then the difference in private shocks has a cumulative distribution function following $\Pr\left(\epsilon_{it}(-1) - \epsilon_{it}(1) < x\right) = \Phi\left(\frac{x}{\sqrt{2}}\right)$, and (11) becomes:

$$\rho^* = \Phi\left(\frac{d}{\sqrt{2}} - \frac{2J\delta(1-2\rho^*)}{\sqrt{2}(2-\delta)}\right) \tag{12}$$

We say that a bad norm persists when $\rho^* \approx 0$ is a possible equilibrium during a sufficiently long period of time in which choice $\omega_i = -1$ is generally preferable from a group welfare perspective. As long as d is large enough relative to J, the only sustainable long-run equilibrium in the system is

the 'good' norm $\rho^* \approx 1$. However, when d is small relative to J so that identity is relatively more important than individualistic returns, two stable equilibria emerge: $p^* \approx 0$ and $p^* \approx 1$. By way of an explicit example, for d-values from 0 to 4, the minimum value of J for which a bad norm of $\rho^* \approx 0$ and a good norm of $\rho^* \approx 1$ can persist is shown in Figure 1.

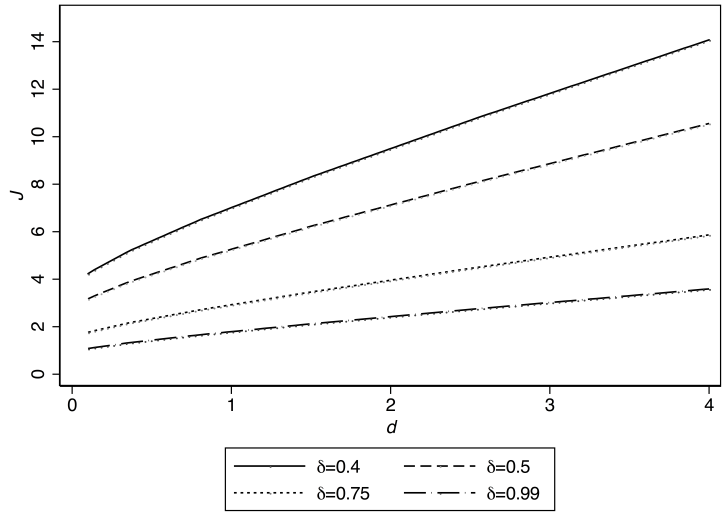


Figure 1: Minimum theoretical J required for bad and good norm coexistence.

Notes: Values are numerically calculated from equation 12 with d-intervals of 0.001. Individuals are assumed to follow a homogeneous threshold rule based on equally weighting their private values and group norm expectations with the form of (9) with different weighting parameters δ. Private shocks for each choice and individual are distributed $\sim \mathcal{N}(0, 1)$.

Continuing the example, consider the parameter space $d = 2$, $J = 8$, and the normal shock distribution described above. Figure 1 shows that for $\delta = 0.5$, a bad norm $\rho^* \approx 0$ can persist. To investigate the likelihood of this occurring, this system was simulated for a group of 100 individuals with self-fulfilling expectations m_i^e following the form of (9) and $\delta = 0.5$. The initial proportion ρ_0 choosing $\omega_i = -1$ was uniformly distributed over $[0, 1]$ and, for each starting value, the game was played for 50 rounds. From 100,000 simulations the bad norm persisted approximately 20% of the time, requiring less than a quarter of the population initially choosing $\omega_i = -1$. Figure 2 shows the result of these simulations. When J is reduced below the persistence threshold to 4, the system stabilizes at $\rho^* \approx 1$ in every simulation; the group always switches to the good norm after 50 rounds. If d_t is allowed to vary slightly around a mean of 2, the results generally hold. Bad norms are now less likely to exist for $J = 8$, but this reduction comes solely from initial values around $\rho_0 = 0.25$; the results are unchanged for initial proportions close to 0.

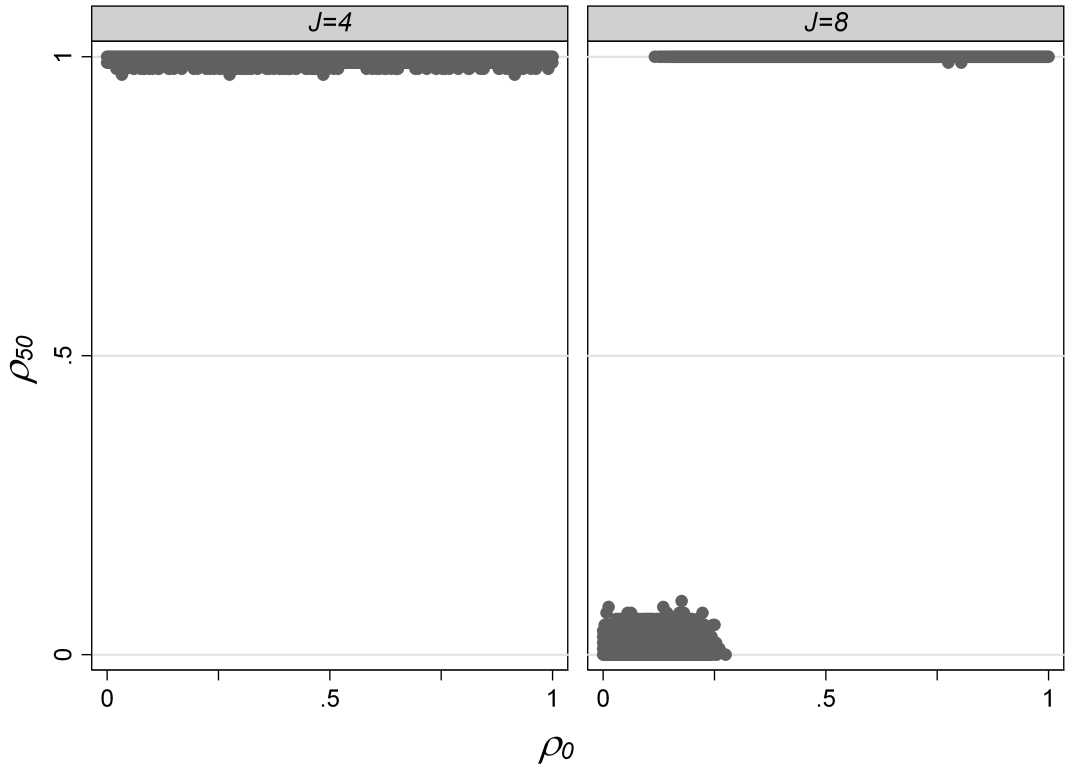


Figure 2: Simulated equilibria for $d = 2, J = 4, N = 100, \delta = 0.5$.

Notes: ρ_t gives the proportion of individuals choosing $\omega_i = -1$ in a round t. Starting proportions are taken from $\sim \mathcal{U}(0, 1)$ across 100,000 simulations of 50 rounds. Individuals are assumed to have expectations of the form specified in (9) with $\delta = 0.5$.

Group size

Following on from (12), the predicted equilibrium proportion, taking each round in isolation, is unaffected. However, in a dynamic model the effects of group size on the persistence of a bad norm manifest themselves more subtly. The probability that at least one group member chooses $\omega_{it} = -1$ increases with N, and so we would expect a higher proportion of rounds with $\rho_{it} > 0$ in larger groups while the bad norm persists. However, the marginal effect of a group member choosing $\omega_{it} = -1$ (a 'deviation' from the norm) on the overall group proportion ρ_{it} is greater for smaller groups.

How do these conflicting forces affect the overall persistence of the bad norm? It can be shown that when a bad norm is in effect, smaller groups are *generally* more likely to reach the tipping proportion in a given round (see Appendix A). This is a consequence of it being less feasible in larger groups that a sufficient proportion of individuals receive extreme shock values in the same round, such that the tipping proportion is breached. The magnitude of size effects is relatively meagre; for small ρ_t and some larger tipping proportion, as might be expected, size differences are

approximated from deep into the tails of a normal cumulative distribution and so the probability of breaching the tipping proportion in a given round approaches zero for all sizes. Further, if there exists some positive probability of reaching the tipping proportion in a given round, a bad norm will eventually be broken over a long enough time horizon. However, over a finite period of multiple rounds these probabilities compound and so some tangible short-term effects may be deduced. For small values of N, the model thus predicts that smaller groups are slightly more likely to switch away from a bad norm, and would be expected to do so faster, than larger groups. Between groups of very large sizes, however, the effect of N on bad norm persistence becomes negligible.

Full information

A final extension to our model with relevance to the laboratory experiment is to consider the dynamic consequences in which individuals know the common values as well as the distribution of private shocks. This may affect individuals' expectations about group behavior. Individuals form these expectations on the basis of the common values, rather than their own private values, and moreover, the certainty provided by information about common utility logically prompts more weighting on this component of expectations formation function. Let δ' represent the weighting parameter for an individual, who otherwise forms expectations with δ, in the presence of full information. A corollary from the function assumed in (9) is then $\psi(d_{it}, m_t) = \delta' m_{t-1} - (1-\delta')\frac{d}{2J}, \quad \delta' \in [0,1]$, where it is assumed that $\delta' < \delta$. Corresponding to (10) of Proposition 2 above, it follows that the equilibrium condition for a stable average group choice becomes:

$$m^* = 2F\left(2J\delta' m^* - (2-\delta')d\right) - 1 \tag{13}$$

The effect of full information on bad norm persistence is ambiguous when contrasted with respect to the previous analysis. The substitution of the common values for an individual's private values in the expectations formation function increases the scope for the 'bad' equilibrium to emerge, while the lower weighting parameter has the opposite effect. In terms of sensitivity, persistence is extremely responsive to changes in δ; a very small decrease in an individual's weighting of the existing group norm causes a large reduction in the scope of bad norm persistence for a given $\{d, J\}$ parameter space[7].

The model provides a testable framework for the role of identity in perpetuating bad norms.

[7] With the benefit of hindsight afforded by our experimental results, an argument could be made that the absence of uncertainty over the common values significantly removes the reliance on historical norms for an individual's expectation about future group behavior. In the extreme case in which $\delta' = 0$, the equilibrium condition reduces to $m^* = 1 - 2F(2d)$, which gives only one 'good' equilibrium for given common values and no longer depends on the identity factor at all.

While it follows that stronger social identity is more likely to foster bad equilibria, the precise conditions under which a bad norm can persist are not trivial. In the absence of convincing behavioral arguments, the weighting parameter $\delta = 0.5$ was arbitrarily chosen for the above simulations of Figure 2. Different weightings produce significantly different results. While (12) cannot be solved analytically, it is clear that individuals must place sufficient weight on the existing norm, relative to the ratio of private and social value considerations, in forming their beliefs in order for a bad norm to persist. For the example above with $d = 2$, a slightly lower weighting parameter of $\delta = 0.4$ would result in $J = 8$ no longer being sufficient for a bad norm to persist; for $\delta = 0.75$, on the other hand, bad norms can now persist for the weaker identity strength of $J = 4$. A laboratory experiment is an appropriate medium through which to investigate these effects further.

3 Experimental design

The computerized experiment was run at the CREED laboratory of the University of Amsterdam. Subjects read the instructions of the experiment at their own pace and had to successful answer some control questions before they could proceed to the experiment[8]. In the experiment, subjects earned points that were converted at the end of each session at an exchange rate of five points for one euro cent (500 points = 1 euro). At the start of the experiment, each subject was randomly assigned to a group and participated in 50 rounds of the Identity Game. Subjects were not told how many rounds the game would last. Points were summed over the 50 rounds and the final game earnings were paid privately. In addition, subjects received a show-up fee of 3 euros.

Recruited was conducted at the University of Amsterdam. Subjects had no prior experience in directly related experiments, and each subject participated in only one session of the experiment. Each session took approximately one hour. Multiple groups were run in each session, but the composition of the groups themselves remained constant. In total, 322 subjects participated in 17 sessions, and earned on average 14.50 euros (s.d. 2.43), including the show-up fee.

The Identity Game used in the experiment featured 50 rounds of the stage game of the model described in the previous Section, but presented in a more subject-friendly manner. In each round players made an individual choice between two 'doors', A and B, from which they could earn points. An individual's payoff depended both on her *private value* and her *social value*. Each door's private value, which an individual observed before making the choice, consisted of the sum of that door's common value and an individual shock. Group members could not observe the components of their private values, but they knew both that the common values were the same for all group members in a given round, and that all shocks were randomly drawn from a standard normal distribution.

[8] Appendix B lists the instructions for the treatment with $N = 6, J = 4$ ("*SmallWeak*"), as well as for the Communication and Full Information treatments. Instructions for the other main treatments differed from *SmallWeak* only with respect to the parameter values.

Social value was determined by the proportion of other group members who made the same choice as an individual, scaled by an identity factor; if an individual was in the minority, the social value was negative. Specifically, the social value to a participant was formulated to a subject in terms of the number of points she would gain (lose) for each other group member who made the same (different) choice as her in a given round.

After the choices by all subjects were submitted in a given round, the payoffs were presented along with information about the number of other group members who chose each door. The experiment then continued to the next round, with subjects next seeing their new private values for the doors.

The common door values used in the experiments were randomly generated in order to create appropriate conditions for testing bad norms and to coincide with the theoretical analysis and simulations. Figure 3 shows how the common door values developed over time in each group of each treatment. Specifically, unknown to the subjects,

- Door A was initially preferred by a large margin (roughly 6 points)
- Common values of each door could change by a maximum of 1 point in each new round
- Door A remained preferable until round 25, after which Door B overtook Door A
- From round 40 until the end of the session, Door B held a positive difference over Door A.

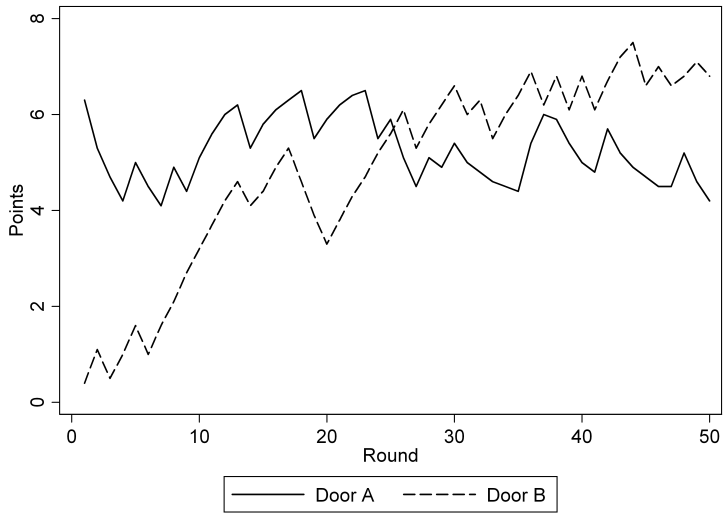


Figure 3: Common door values

Notes: For participants in the laboratory experiment, all values were multiplied by 10.

These stipulations were designed to create an environment in the first half of the session in which

Table 1: Treatments

Treatment	Identity factor J	Group size N	#Groups	Bulletin board?	Common info?
SmallWeak	4	6	8	No	No
SmallStrong	8	6	8	No	No
BigWeak	4	11	7	No	No
BigStrong	8	11	7	No	No
Communication	8	6	8	Yes	No
Full information	8	6	4	No	Yes

a social norm of choosing Door A could emerge, which, after 25 rounds, would then be consistently the socially inefficient choice.

The linear nature of the social value was explained in terms of the number of points earned per other player making the same choice; for example, in the treatment with $N = 6$ and $J = 4$ ("*SmallWeak*"), the instructions contained the sentence:

> *You **gain 8 points** for every person who makes the **same** choice as you, but you **lose 8 points** for every person who makes the **opposite** choice to you.*

Notice that in the experiment, like in the theoretical model, an individual i thus receives a payoff according to equation 1 in round t, where $\omega_{it} = 1$ is defined as choosing Door A, $\omega_{it} = -1$ as choosing Door B, m_{it} as the average choice of the others in the group, and J as the identity factor.

All treatments made use of the experimental variant of the Identity Game described above. Table 1 summarizes the main features of the treatments. These were varied between subjects, with the four main treatments based on combinations of the two parameters of interest: identity strength and group size. We discuss two additional treatments after we have explained the main treatments.

Private shocks were randomly drawn from $\sim \mathcal{N}(0, 1)$ for each individual, door and round. Realizations of private shock distributions for each individual were matched for treatments with the same group size. That is, each of the 8 groups in *SmallWeak* had a matched group in *SmallStrong* with the same private shocks distributed across group members, doors and rounds, and likewise for the 7 groups in each of the larger treatments. All generated private shocks and common values were multiplied by 10 and rounded to the closest integer to avoid decimal points.

The group sizes were chosen to make it easier for subjects to calculate the potential social values, which required considering fractions of 5 or 10. The identity factors were chosen to coincide with the theoretical simulations predicting mixed results when subjects assign equal weights to both

the existing norm and their own private information in forming their expectations ($\delta = 0.5$). With these weights, the model predicts that groups will initially coordinate on Door A, which becomes the group norm, and are more likely to switch to Door B by round 50 when identity is weak. From the dynamic analysis it follows that the effect of group size after 50 rounds is theorized to be relatively small, with any differences likely to manifest themselves by groups of smaller size switching to Door B faster, if at all. This yields two specific hypotheses about the group proportions after 50 rounds:

Hypothesis 1. *Groups are more likely to stay with choosing Door A after it has become the bad norm when $J = 8$ than when $J = 4$.*

Hypothesis 2. *Groups are equally likely to stay with choosing Door A after it has become the bad norm when $N = 6$ or 11.*

We extended the experiment to include two additional treatments that share a common theme of reducing the subjective uncertainty about group behavior. Unlike in the the main treatments, subjects were offered the possibility to communicate in the *Communication* treatment. In every round before they chose their door, each subject could express her intention on a 'Bulletin Board'. Posts on the Bulletin Board were anonymous. Subjects were informed that there was no obligation to honor a post, and that it was also possible not to post anything. After everyone had made their decisions about posting for that round, group-members saw the total number of posts (or 'intentions to choose') for Door A and Door B before they actually made their final choice of door. All other features of this treatment were the same as for *SmallStrong*. The *Communication* treatment was run after the other treatments in order to investigate whether bad norm persistence could be weakened by allowing for anonymous communication.

Hypothesis 3. *Bad norms are more easily broken when there is a possibility to anonymously communicate intended choices.*

Finally, in the small full-information treatment subjects could precisely see the decomposition of their private values into the common values and their own personal shocks for each door in every round (in the other treatments subjects were only informed of the sum; the decomposition was never revealed). Both this 'Full Information' treatment and the Communication treatment can be thought of as reducing the uncertainty pertaining to group payoffs at each choice.

Hypothesis 4. *Bad norms are more easily broken when subjects receive complete information of the common values and their own private shocks.*

In each round of each treatment, subjects' screens displayed the round number, the cumulative earnings, the private values for each door, a choice button for Door A or Door B to be submitted, and a history footer. The history footer contained the total history of the proportion of other group

members making each choice for every completed round[9]. At the end of round 50, subjects filled out a short questionnaire before they were paid.

4 Results

We present the results in two parts. Section 4.1 provides the results of the main treatments. It clarifies the circumstances under which bad norms emerge and persist. Section 4.2 investigates how bad norms can be broken. This section also sheds light on the role that pluralistic ignorance plays in the persistence of bad norms.

4.1 Emergence and persistence of bad norms

Figure 4 displays the frequency of norm breaking by treatment. None of the groups with the strong identity factor ($J = 8$) switched to Door B by round 50, regardless of group size. When the identity factor was weakened to J=4, five out of the eight groups (62.5%) in *SmallWeak* switched to Door A, while three out of seven (42.9%) did the same in the *BigWeak* treatment. The simulations of the theoretical model for the common values, shocks and treatments used in the experiment also produce a slight favoritism for *SmallWeak* compared to *BigWeak* for the sequence of common values used. Calibrating the experimental results to simulations from the model with the same common values and an expectations function of the form of (9) produces $\delta = 0.75$ to give the best fit to the results, suggesting that subjects placed relatively more weight on the group norm than their own private values. The calibration process chose δ to minimize the sum of squared differences between the simulated and experimental proportions across treatments.

Table 2 demonstrates that the descriptive statistics of the data partitioned by treatment are similar when norm breaking is defined by different measures, such as the average ρ across all rounds, the final rounds, or from round 26-50, (the rounds after which the common value of Door B overtakes that of Door A). Detailed proportions for the 30 individual groups can be found in Appendix C. For each individual group, the average group choice stuck closely to the two theoretical stage-game equilibria of $\rho = 0$ and 1 across the rounds; groups spent few rounds in the socially destructive mixed proportions around $\rho = 0.5$. For the groups that finally broke the norm, once approximately a third of the group had simultaneously chosen Door B the group generally took little time in reaching the more favorable equilibrium[10].

The first key result reflects our hypothesis regarding the strength of the identity factor. The upper panel of Table 2 clarifies that identity has a substantial impact on the proportion switching

[9] An example screenshot is displayed in the Appendix.

[10] In the context of the theoretical model, this would suggest a tipping proportion $\tilde{\rho} \approx \frac{1}{3}$.

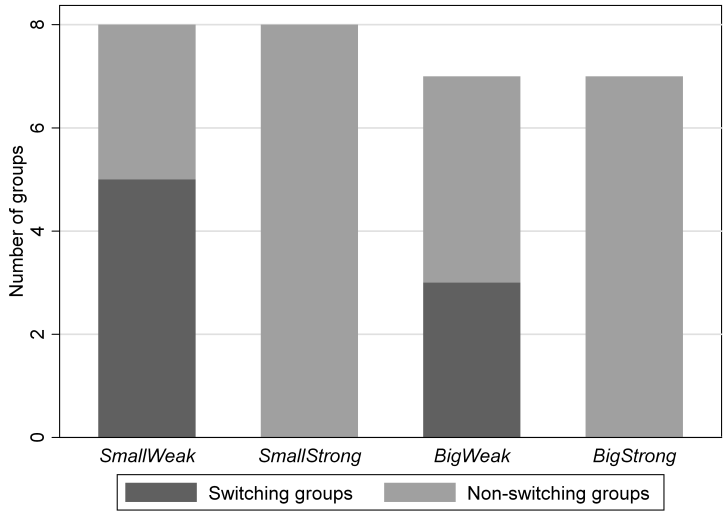


Figure 4: Switching groups by treatment

Notes: 'Switching' is defined as more than half of the group choosing Door B in round 50 ($\rho_{50} > 0.5$).

Table 2: Key performance indicators by treatment

	Treatments	ρ_{50}	$\bar{\rho}_{(45-50)}$	$\bar{\rho}_{(t\geq 26)}$	$\bar{\rho}_{all}$	$\bar{t}_{switch}$
	SmallWeak	.65	.62	.46	.26	29.6
	SmallStrong	.00	.00	.03	.03	-
	BigWeak	.47	.36	.26	.14	39.0
	BigStrong	.00	.02	.02	.02	-
Testing identity:	*SW vs SS*	.00***	.00***	.01***	.01***	
	BW vs BS	.02**	.04**	.11	.06*	
Testing group size:	*SW vs BW*	.46	.41	.30	.30	
	SS vs SS	.12	.02**	.82	.56	

Notes: In the upper panel, values are averages of the group values within each treatment. ρ_{50} is the final group proportion choosing Door A. $\bar{\rho}_{(45-50)}$ is the average ρ across the last final six rounds. $\bar{\rho}_{all}$ is the average ρ across all rounds. $\bar{\rho}_{(t\geq 26)}$ is the average ρ from round 26, when the common value of Door A becomes larger than that of Door B. $\bar{t}_{switch}$ is the average switching time, considering only those groups that switched to Door B by round 50. In the lower panels, p-values are derived from Mann-Whitney rank sum tests. In the tests, each group yields one observation.

to the good door in the latter part of the experiment. When identity is strong, all groups stay with Door A after it has become the bad door. The lower panels of Table 2 show the extent to which the results differ systematically across treatments. An increase in identity significantly enhances various measures of ρ for both $N = 6$ and $N = 11$.

RESULT 1: *Bad norms are more likely to persist when group identity is strong.*

The result is further illustrated in Figure 5. Only groups with the smaller identity factor switched their overall door preference after round 25. The figure also reveals that groups that switched to Door B of size $N = 6$ generally did so earlier than the switching groups of size $N = 11$, although these short-run size effects disappeared by the end of the 50 rounds.

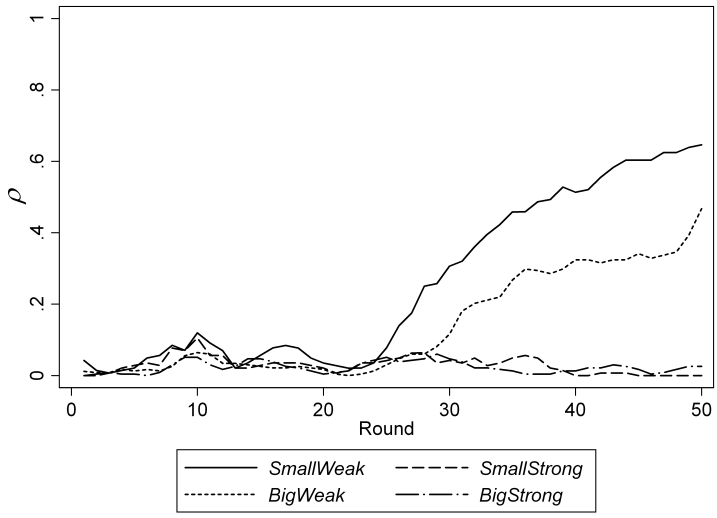


Figure 5: Average round-by-round group choice by treatment

Notes: Each treatment line depicts the average group proportion choosing Door B across all groups in the treatment. Lines have been smoothed via a three-round equally weighted moving average.

The second key result concerns the role of group size. This has a much smaller effect on the emergence and persistence of the bad norms. The tests on group size reported in the lower panel of Table 2 tend to be insignificant. When only the weaker identity groups are considered, the graphical representation of round-by-round pooled data presented in Figure 5, when broken by group size, does suggest faster deviations from the norm for $N = 6$. However, it is conceivable that the two lines would have converged if the experiment had been extended beyond 50 rounds, so it is impossible to claim a long term group size effect on the eventual persistence or collapse of bad norms.

RESULT 2: *The persistence of bad norms does not depend on group size in the long run.*

Nevertheless, in the *short term*, there is some evidence that individuals are less willing to go against the norm when within larger groups. In the first 20 periods, for example, although the common value of Door A was always preferred, some individuals received private shocks such that there was an individual incentive to deviate from the norm. Subjects were significantly more likely to deviate when group size was smaller, as evidenced from rank-sum tests of the averaged ρ of rounds 1-20 (*J=4:* Mann-Whitney $p = 0.03$; *J=8:* Mann-Whitney $p = 0.02$).

This generates support for the mechanism predicted by the model to cause some short-run size effects. Holding identity strength and other parameters constant, the model predicts that, while the bad norm persists, larger groups would more frequently experience rounds with at least one person deviating, but that these rounds would on average have a lower ρ. Table 3 shows that when we control for J, the experimental results confirm these predictions.

Table 3: Deviation statistics during bad norm persistence by treatment

	SmallWeak	*SmallStrong*	*BigWeak*	*BigStrong*
Frequency of deviation rounds	27.5%	15.8%	34.7%	21.7%
Average ρ in deviation rounds	0.191	0.183	0.143	0.109

Notes: Values are averages of the group values within each treatment, restricted to rounds of bad norm persistence ($\rho < 0.5$). Frequency of deviation rounds is calculated by dividing the number of rounds with deviations by the total number of rounds with $\rho < 0.5$.

Interestingly, for $\bar{\rho}_{(1-20)}$ the measure generating tangible short-run size effects, identity strength, was not found to be significant. It can be gleaned that in these early rounds when Door A is still commonly preferable, it is the size of the group, rather than identity, that determines subjects' predilection to deviate for individual reasons. However, the severity of the loss that usually follows for a subject who decides to deviate depends on the identity factor (manifested in the social value). This severity then determines the likelihood that the individual returns to the group choice or continues to deviate in the subsequent round. To sum up, the evidence suggests that identity strength is chiefly responsible for whether a bad norm persists, while group size plays a role in the short term and in determining the speed of a norm shift[11].

[11]Groups that do not stay with the bad norm appear to benefit from the presence of 'Leaders'. Leaders are defined as individuals who choose Door B in two consecutive rounds $t, t+1$ when $\rho_{t-1}, \rho_t < 0.5$. They may be thought of as sacrificing personal gain in order to signal the group and put pressure on the norm, and their presence is highly correlated with breaking down the norm. None of the ten groups in which no Leader emerged managed to switch to Door B. Whether the presence of Leaders is in itself conducive to collapsing a bad norm is an open question, as clear endogeneity issues are present. However, controlling for identity, there is a strong positive correlation between the proportion of Leaders in a group and the collapse of the bad norm. The difference in the percentage of Leaders for groups that persist with choosing Door A or eventually switch to Door B is highly significant (Mann-Whitney p=0.01).

The enduring social welfare inefficiency of groups that persist with the bad norm is somewhat reflective of situations with *pluralistic ignorance*. As discussed in the introduction, pluralistic ignorance is a phenomenon whereby most individuals in a group have a positive personal incentive to deviate from the norm, but believe that the majority of group members have a private incentive to keep to the status quo. In this experiment, beliefs causing pluralistic ignorance can be considered to have been incorporated into the social welfare function by way of expectations outweighing own private value considerations.

If all individuals in a group have a private value of Door B exceeding that of Door A in a particular round of the experiment, but *all* group members choose Door A ($\rho = 0$), the group is said to exhibit *total pluralistic ignorance*. Such incidence represents the worst case scenario from a social welfare perspective; in fact, if social value was ignored, any other proportion of choices would be a Pareto improvement. In the experiment the number of rounds in which total pluralistic ignorance could potentially exist is naturally higher for smaller groups, as groups with more individuals are more likely to produce at least one group member realizing extreme private shocks. Figure 6 compares the number of potential rounds of total pluralistic ignorance to those that eventuated in the experiment. This again reveals a strong identity effect. *SmallStrong* and *BigStrong* saw total pluralistic ignorance in, respectively, an average of 87% and 81% of each treatment's potential rounds, while for *SmallWeak* and *BigWeak* the average frequencies were 27% and 31%.

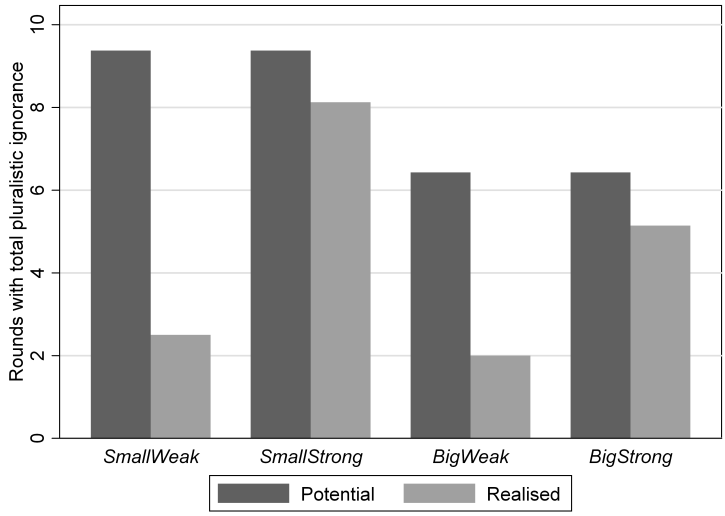


Figure 6: Mean potential and realized rounds of total pluralistic ignorance.

Notes: A 'total pluralistic ignorance' round is defined as a round t in which all players receive $d_{it} > 0$ and subsequently choose Door A ($\rho_t = 0$). Amounts are averages per group out of a total of 50 rounds.

4.2 Breaking bad norms and preventing their emergence

A natural extension to our setup is to introduce communication for the participants. Recent examples of the much-heralded role of the internet in eroding sexual discrimination in India and inciting revolutionary action in Egypt add some weight to the role of communication in breaking down historically powerful social norms. Online social media facilitates cost-free, anonymous communication to a wide audience, allowing individuals with a private interest in changing the status quo to signal their desire for change in a broad manner. We symbolized these opportunities in a further treatment in which participants were given the option in each round to indicate their choice intentions. Subjects could choose one of two posts to an anonymous 'Bulletin Board' - "I intend to choose Door A" or "I intend to choose Door B" - or not to post at all.

We replicated the *SmallStrong* treatment by running eight groups with $N = 6$ and $J = 8$ (48 subjects), which, in the original treatment, produced no norm breakages. With the addition of anonymous 'cheap talk', however, all eight groups easily managed to break the bad norm[12]. Only two of the 48 participants chose not to use the Bulletin Board at all; of the rest, most subjects took the opportunity to post in every round. Moreover, the collection of posts on the Bulletin Board was overwhelmingly indicated as the primary means of expectation formation in the answers to the questionnaire. Figure 7 presents the average number of announcements to opt for Door B together with the actual choices for Door B as the rounds unfolded. For all eight groups, the switch in average group indications from Door A to Door B coincided with the shift in the difference in common values. Interestingly, all participants exploited the anonymity by acting contrary to their posted indication in at least one round (mean = 5.5 rounds, s.d. = 2.2). This fact, in addition to the absolute switch in results in comparison to *SmallStrong*, suggests some natural extensions. It would be of interest to see how subjects react when the veil of anonymity is removed, or if communication opportunities are limited either to less regular intervals or to only a subset of the population.

It is clear from the results of this additional treatment that communication can play a significant role in assisting in the breakdown of a bad norm. We believe a natural explanation for these results is that the 'cheap talk' may serve to reduce ambiguity about future social utility. This motivates the question: how is group behavior affected when communication is prohibited but individuals are made aware of the expected payoffs of their fellow group members? We ran four groups with parameters $N = 6$, $J = 8$, with the only difference to the *SmallStrong* treatment of the main experiment being that subjects could precisely see the common values and their own shocks for each door in every round. Figure 8 shows that with full information over the decomposition of the

[12]In the first of two sessions, a programming bug incorrectly displayed the earnings total as twice the actual earnings. Participants were still able to deduce their cumulative earnings from the round-by-round earnings, which were displayed correctly. In the second session, subjects were informed of the display error and given a calculator in case they wished to calculate their cumulative earnings from the round-by-round displays; group results were similar across both sessions.

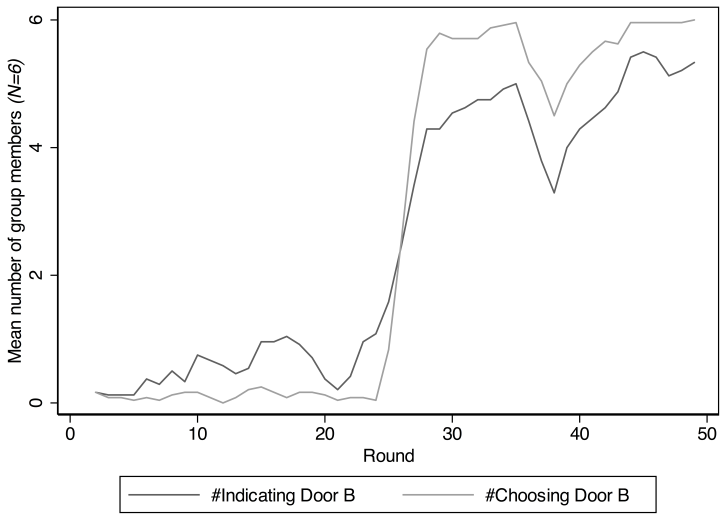


Figure 7: Average round-by-round group indications and actual choices of "Door B" for the communication treatment

Notes: Treatment parameters were: $N = 6$, $J = 8$. Almost all subjects in a group posted their intentions in every round (mean = 5.6, s.d. = 0.6). Lines have been smoothed via a three-round equally weighted moving average.

private values for each individual, groups broke the norm only slightly later than in the treatment where they could communicate. In the long run they were as successful in deviating from the bad norm as in the Communication treatment. This accords with psychological theories of social norms that propose that payoff uncertainty of other group members is a crucial ingredient for bad norm persistence[13]. In both the Communication and Full Information treatments, no group ever exhibited total pluralistic ignorance, as defined above, for any round[14].

[13] E.g. Sherif (1936)

[14] Note that the final treatment is referred to as 'Full Information' because each individual can see for each round the common values and their *own* private shocks, which allow them to form a true expectation about the private values of the other group members. Individuals do not, however, know the precise realizations of the private shocks for the others.

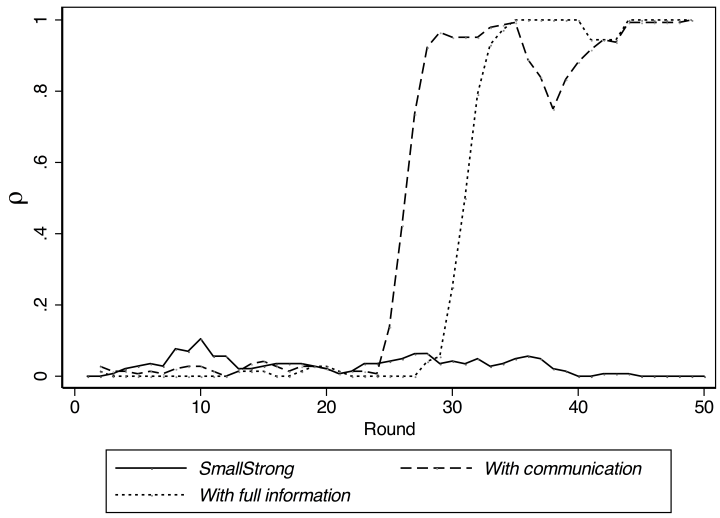


Figure 8: Average round-by-round group choice for $N = 6, J = 8$, including anonymous communication and decomposed private values (full information) treatments

Notes: Each treatment line depicts the average group proportion choosing Door B across all groups in the treatment. Lines have been smoothed via a three-round equally weighted moving average.

5 Discussion

The experimental results confirm the fundamental prediction of the theoretical model: Bad norms can persist in the laboratory when group identity is strong relative to the difference in private payoffs. Bad norms emerge as a result of a good equilibrium gradually becoming a bad equilibrium in a coordination structure due to changing payoffs over time. Once established, these bad norms can persist so long as the personal incentives to deviate are small and social identity is strong. Smaller groups have a better chance of collectively breaking a bad norm in the short term, but over a longer horizon the prospects between differently sized groups even out.

The results support the modest short-term size effect predicted by the model, although its magnitude was more pronounced in the laboratory than when simulated. Although not statistically significant, this stronger effect of group size in the short run coincides with the findings from some conformity experiments in social psychology. An explanation for groups of smaller sizes exhibiting more pronounced switching behavior could be found in the well-known 'bystander effect', which speaks to the drawbacks of increased diffusion of responsibility in larger groups. Such psychological effects regarding group size, particularly with regard to incentives to 'lead' the group out of a bad equilibrium, are not captured by our model. In the early rounds of the experiment, while a good social norm of choosing Door A was in place, individuals were found to be more likely to deviate on account of private incentives in the smaller groups. One psychological explanation for this result may be that individuals feel a sense of persuasive power and influence in smaller groups above that

which is implied from the social value function. The effect of group size on this influential self-belief is worthy of further inspection.

Given the short-term size effects revealed from both the theoretical and experimental results, the time horizon for repeatedly considering a choice deserves reflection. In particular, there are many social norms in the real world that preside over environments in which individual decisions are made infrequently, or for which the consequences of a certain choice may be irreversible. A woman's decision to undergo genital mutilation is not encountered often in a lifetime and, once chosen for, is normally irreparable. The decision to rebel against an ruthlessly oppressive government is a choice that may have permanent (possibly fatal) consequences with no further opportunity of revision. In such circumstances without the regularity of repeated decisions, the likelihood of a tipping proportion of individuals being simultaneously personally incentivized to deviate from the status quo may be very low. Bad social norms affecting infrequent and potentially irrevocable choices of this nature may thus display even higher levels of persistence.

An important insight from our experiment is that strong feelings of group identity are a necessary but not sufficient condition for the persistence of bad norms. That is, when strong feelings of group identity are paired with full information about the preferences of others, bad norms disappear. A similar beneficial effect results from communication. We reason from our empirical findings that an important condition for bad norm persistence is ambiguity about other group members' incentives and future behavior. Our results motivate a need for further tests in the field, and suggest that bad norm interventions that target ambiguity may be worthy of consideration.

Finally, it should be stated that the debate over the true effect of social identity has not reached a consensus in nearly a century of academic investigation. The experimental design automatically monetizes identity effects into individuals' payoffs, but further research could consider directly triggering group identity in the laboratory. What a more natural setting of this nature loses in robustness would be compensated by adding support to the behavioral foundations of the modeling of bad social norms proposed in this paper.

6 Appendix A: Effect of group size

Consider a scenario in which the bad norm $\omega_{it} = 1$ is persistent on account of relatively large J and m^e_{it}, such that in the majority of rounds $\rho_{it} = 0$. *Ex ante*, the probability of an individual choosing $\omega_{it} = -1$ in a given round t is $\hat{\rho}_t$, regardless of the group size. Now consider the rounds in which $0 < \rho_{it} < 0.5$; that is, the bad norm $\omega_i = 1$ is still in effect but *at least one* group member receives a private shock difference large enough to induce choosing $\omega_{it} = -1$. This likelihood is not the same across group sizes. The probability that at least one group member chooses $\omega_{it} = -1$ increases with N, and so we would expect *a higher proportion of rounds with* $\rho_{it} \neq 0$ *in larger groups* while the bad norm persists. However, the marginal effect of a group member choosing $\omega_{it} = -1$ on the overall group proportion ρ_{it} decreases with N, and so of those rounds where $\rho_{it} \neq 0$ while the bad norm persists, we would expect that ρ_{it} *is higher on average for smaller groups.*

Now, assume there is some 'tipping proportion' $\tilde{\rho}$ that, if reached after a previous equilibrium of full conformity to the bad norm ($\rho^* \approx 0$), would result in a switch to the 'good' equilibrium $\rho^* \approx 1$ with almost certainty. The tipping proportion is greater than the predicted group proportion $\hat{\rho}_t$ so that on expectation it should not be breached in a given round. Then, after a round in which $\rho_{t-1} \approx 0$, the probability of reaching the tipping proportion in round t is the probability that at least $N\tilde{\rho}$ individuals choose $\omega_{it} = -1$. From the researcher's perspective, the number of individuals choosing $\omega_{it} = -1$ follows a binomial distribution so that $N\rho_t \sim \mathcal{B}(N, \hat{\rho}_t)$ and hence:

$$\begin{aligned} \Pr(\rho_t \geq \tilde{\rho}) &= 1 - \Pr(\rho_t < \tilde{\rho}) \\ &= 1 - \sum_{j=0}^{\lfloor N\tilde{\rho} \rfloor} \binom{N}{j} \hat{\rho}_t^j (1 - \hat{\rho}_t)^{N-j} \end{aligned} \tag{14}$$

where $\lfloor N\tilde{\rho} \rfloor$ is the largest integer less than $N\tilde{\rho}$.

This function does not change monotonically with N. However, some idea can be garnered as to how the probability is affected across general size increases. The binomial distribution can be approximated by a normal distribution with mean $N\hat{\rho}_t$ and variance $N\hat{\rho}_t(1 - \hat{\rho}_t)$ when $N\hat{\rho}_t > 5$. Assuming this is met, equation (14) can be approximated by:

$$\begin{aligned} \Pr(\rho_t \geq \tilde{\rho}) &= 1 - \Pr\left(\frac{N(\rho_t - \hat{\rho}_t)}{\sqrt{N\hat{\rho}_t(1 - \hat{\rho}_t)}} < \frac{N(\tilde{\rho} - \hat{\rho}_t)}{\sqrt{N\hat{\rho}_t(1 - \hat{\rho}_t)}} \right) \\ &\approx 1 - \Phi\left(\sqrt{N} \frac{\tilde{\rho} - \hat{\rho}_t}{\sqrt{\hat{\rho}_t(1 - \hat{\rho}_t)}} \right) \end{aligned} \tag{15}$$

which, for $\tilde{\rho} > \hat{\rho}_t$, is a decreasing function of N.

When a bad norm is in effect, smaller groups are thus *generally* more likely to breach the tipping proportion in a given round. The effect of size on persistence increases slowly and not monotonically, although comparisons can be made for sizes that are not very close together. This is due to the discrete nature of the possible proportions and hence the upper sum limit $\lfloor N\tilde{\rho} \rfloor$.

7 Appendix B: Instructions and screenshot

7.1 Instructions for *SmallWeak* (*N=6, J=4*)

Welcome to this experiment on decision-making. Please read the following instructions carefully. When everyone has finished reading the instructions and before the experiment starts, you will receive a handout with a summary of the instructions. At the start of the experiment, you will be randomly assigned to a group of 6 participants. Throughout the experiment you will stay in the same group. You will play a number of rounds (at least 30, but not more than 80) in which you will make decisions. In the experiment, you will receive a starting capital of 1500 points. In addition, you earn and sometimes lose points with your decisions in the rounds. These amounts will be added to (or subtracted from) your starting capital. At the end of the experiment, your final point earnings will be exchanged for euros. Five points will be exchanged for 1 eurocent. Therefore 500 points will earn one euro.

Each round, every participant in the group will make a decision between "Door A" and "Door B". The payoff you receive from choosing a particular door in a round will be the sum of two parts, based on:

- Your **private value** for the door (which could be positive, zero or negative), and
- Your **social value** for the door (which could also be positive, zero or negative).

Private value

At the start of each round, you will be informed of your own private value for each door. Private values are generated as follows: At the start of a round, we will draw **common values** for each door, which no subject can see and which may change in each new round. The common value for a door will be the same for every participant in your group. However, the two doors will most often have different common values. For each door, we will then draw **individual shocks** for each participant, which again no subject can see. For each door, every participant's private shock is randomly drawn from a normal distribution (with an average value of 0 and a standard deviation of 10). The graph below clarifies how frequently different private shocks occur.

Each participant receives an independent private shock for each door. Therefore, the private shocks for one participant usually differ from the private shocks of the other participants. We then add the common value for each door to your private shock for that door, which gives you your **private value**. Therefore, for each door, your private value could be higher or lower than the average private value of your group. No other participant can see your private values.

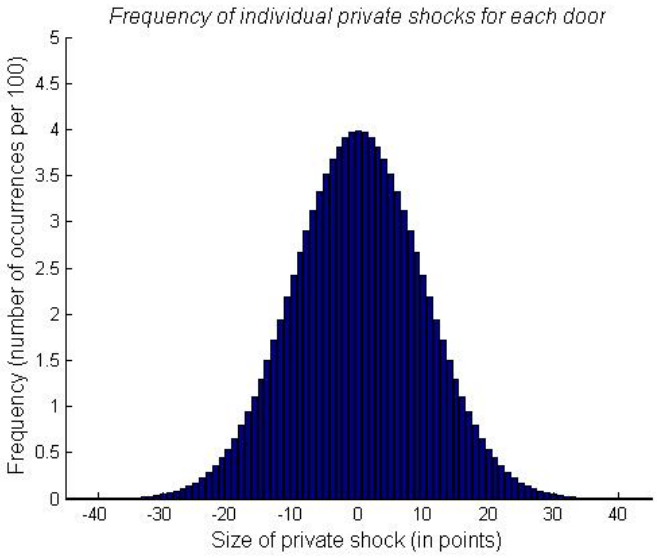


Figure 9

Social value

Your social value in a round depends on how many other people in your group make the same door choice as you. You gain if the majority of the other participants make the same choice as you, but you make a loss if the majority makes the other choice. Specifically, you **gain 8 points** for every person who makes the **same** choice as you, but you **lose 8 points** for every person who makes the **opposite** choice to you. As there are five other people in your group, you can get a maximum social value of 40 points if everyone chooses the same door as you, or you can maximally lose 40 points if everyone chooses the other door to you.

The other participants in your group face the same decision as you do. That is, they receive similar information as you do (although their private values will most likely differ), they also choose between Door A and Door B and they make money in the same way as you do.

Example

In this game, there are 5 other participants in your group. So, for example, if you choose Door A with a private value of 60 points and 4 others also choose Door A, your payoff equals your **private value** (60) plus a **social value** (32 - 8 = 24), for a **total of 84 points**.

If on the other hand you choose Door B with a private value of 50 points and the 5 others choose Door A, your payoff equals your **private value** (50) *minus* a **social value** of 40 points, for a **total of 10 points**.

Sequence of events

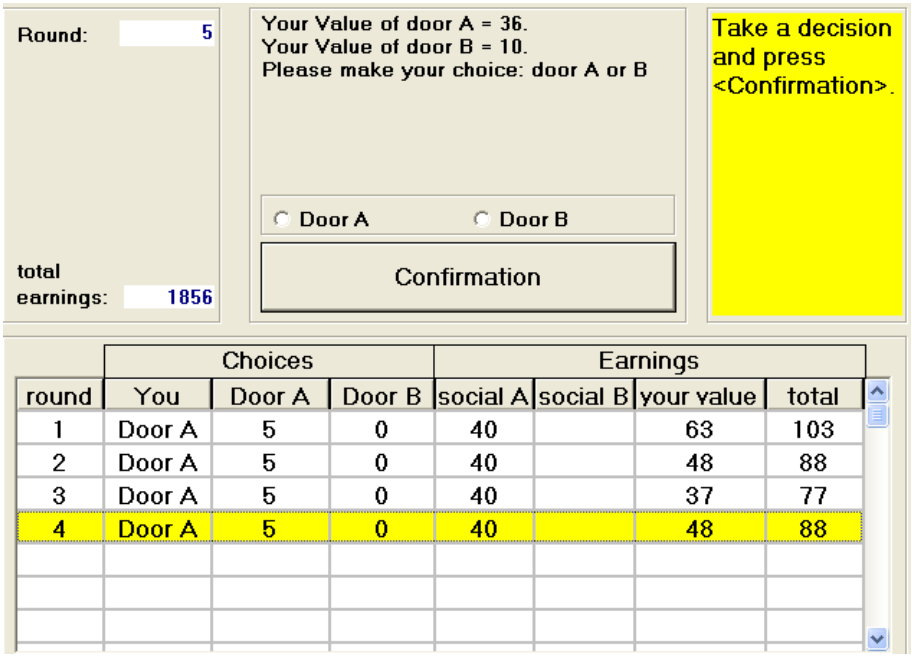
Summing up, each round is characterised by this sequence of events:

- At the start of each round, you are told your private values for the doors.
- You make your choice between Door A and Door B.
- At the end of a round, you are told the number of your group members who made each choice, what the social values were for those who chose each door, and you are informed of your payoff in that round. Each round's payoff is the sum of your chosen door's **private value** and your chosen door's **social value**.

Other participants face exactly the same sequence of events.

You can always see the history of the groups choices for all rounds up to that point at the bottom of your screen. You can also always see the sum of the number of points that you earned so far at the top left corner of your screen.

On the next screen you will be requested to answer some control questions. Please answer these questions now.



	Choices			Earnings			
round	You	Door A	Door B	social A	social B	your value	total
1	Door A	5	0	40		63	103
2	Door A	5	0	40		48	88
3	Door A	5	0	40		37	77
4	Door A	5	0	40		48	88

Figure 10: Screenshot of individual in *SmallWeak* treatment

Notes: Screenshot is taken from the start of round 5. The history footer has a scroll function such that the complete history up until the current round is accessible. Theoretical values were multiplied by 10 in the experiment.

7.2 Instructions for *Communication*

Welcome to this experiment on decision-making. Please read the following instructions carefully. When everyone has finished reading the instructions and before the experiment starts, you will receive a handout with a summary of the instructions. At the start of the experiment, you will be randomly assigned to a group of 6 participants. Throughout the experiment you will stay in the same group. You will play a number of rounds (at least 30, but not more than 80) in which you will make decisions. In the experiment, you will receive a starting capital of 1500 points. In addition, you earn and sometimes lose points with your decisions in the rounds. These amounts will be added to (or subtracted from) your starting capital. At the end of the experiment, your final point earnings will be exchanged for euros. Five points will be exchanged for 1 eurocent. Therefore 500 points will earn one euro.

Each round, every participant in the group will make a decision between "Door A" and "Door B". The payoff you receive from choosing a particular door in a round will be the sum of two parts, based on:

- Your **private value** for the door (which could be positive, zero or negative), and
- Your **social value** for the door (which could also be positive, zero or negative).

Private value

At the start of each round, you will be informed of your own private value for each door. Private values are generated as follows: At the start of a round, we will draw **common values** for each door, which no subject can see and which may change in each new round. The common value for a door will be the same for every participant in your group. However, the two doors will most often have different common values. For each door, we will then draw **individual shocks** for each participant, which again no subject can see. For each door, every participant's private shock is randomly drawn from a normal distribution (with an average value of 0 and a standard deviation of 10). The graph below clarifies how frequently different private shocks occur.

Each participant receives an independent private shock for each door. Therefore, the private shocks for one participant usually differ from the private shocks of the other participants. We then add the common value for each door to your private shock for that door, which gives you your **private value**. Therefore, for each door, your private value could be higher or lower than the average private value of your group. No other participant can see your private values.

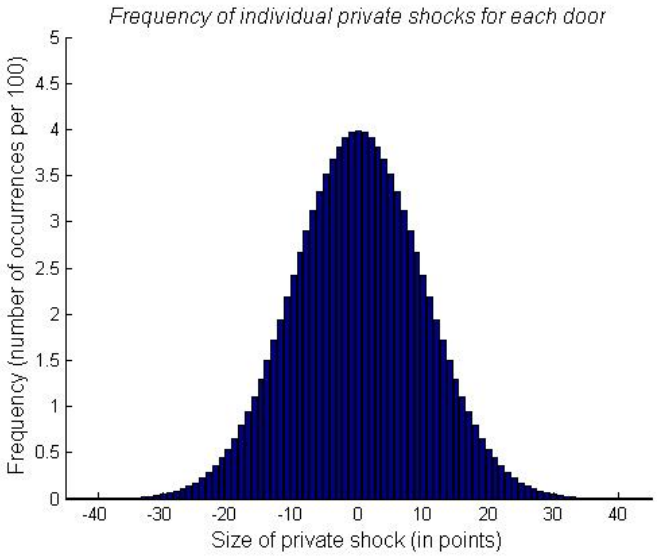


Figure 11

Social value

Your social value in a round depends on how many other people in your group make the same door choice as you. You gain if the majority of the other participants make the same choice as you, but you make a loss if the majority makes the other choice. Specifically, you **gain 8 points** for every person who makes the **same** choice as you, but you **lose 8 points** for every person who makes the **opposite** choice to you. As there are five other people in your group, you can get a maximum social value of 40 points if everyone chooses the same door as you, or you can maximally lose 40 points if everyone chooses the other door to you.

The other participants in your group face the same decision as you do. That is, they receive similar information as you do (although their private values will most likely differ), they also choose between Door A and Door B and they make money in the same way as you do.

Example

In this game, there are 5 other participants in your group. So, for example, if you choose Door A with a private value of 60 points and 4 others also choose Door A, your payoff equals your **private value** (60) plus a **social value** (32 - 8 = 24), for a **total of 84 points**.

If on the other hand you choose Door B with a private value of 50 points and the 5 others choose Door A, your payoff equals your **private value** (50) *minus* a **social value** of 40 points, for a **total of 10 points**.

Bulletin Board

In every round, **before you choose your door**, you can indicate your intentions. On the *Bulletin Board*, which everyone can see, you can choose to post that you intend to choose Door A or Door B. Posts are **anonymous** and there is no obligation to honour your posts. Alternatively, you can also elect not to post anything. After everyone has made their decision about posting for that round, you will be able to see the total number of posts for Door A and Door B on the *Bulletin Board* before finally choosing your door.

Sequence of events

Summing up, each round is characterised by this sequence of events:

- At the start of each round, you are told your private values for the doors.
- You can choose either to anonymously post on the *Bulletin Board*, or not to post at all.
- You see the number of posts for each door on the *Bulletin Board*.
- You make your choice between Door A and Door B.
- At the end of a round, you are told the number of your group members who made each choice, what the social values were for those who chose each door, and you are informed of your payoff in that round. Each round's payoff is the sum of your chosen door's **private value** and your chosen door's **social value**.

Other participants face exactly the same sequence of events.

You can always see the history of the groups choices for all rounds up to that point at the bottom of your screen. You can also always see the sum of the number of points that you earned so far at the top left corner of your screen.

On the next screen you will be requested to answer some control questions. Please answer these questions now.

7.3 Instructions for *Full Information*

Welcome to this experiment on decision-making. Please read the following instructions carefully. When everyone has finished reading the instructions and before the experiment starts, you will receive a handout with a summary of the instructions. At the start of the experiment, you will be randomly assigned to a group of 6 participants. Throughout the experiment you will stay in the same group. You will play a number of rounds (at least 30, but not more than 80) in which you will make decisions. In the experiment, you will receive a starting capital of 1500 points. In addition, you earn and sometimes lose points with your decisions in the rounds. These amounts will be added to (or subtracted from) your starting capital. At the end of the experiment, your final point earnings will be exchanged for euros. Five points will be exchanged for 1 eurocent. Therefore 500 points will earn one euro.

Each round, every participant in the group will make a decision between "Door A" and "Door B". The payoff you receive from choosing a particular door in a round will be the sum of two parts, based on:

- The **common value** of the door (which is the same for all participants),
- Your **private value** for the door (which could be positive, zero or negative), and
- Your **social value** for the door (which could also be positive, zero or negative).

Common value

At the start of a round, you will be told the common value for each door, which everyone can see, and which may change in each new round. The common value for a door will be the same for every participant in your group. However, the two doors will most often have different common values.

Private value

At the start of each round, you will be told your private value for each door, which will be the same for every round and which no other participant can see. For each door, every participant's private value is randomly drawn from a normal distribution (with an average value of 0 and a standard deviation of 10). The graph below clarifies how frequently different private values occur. Each participant receives an independent private value for each door. Therefore, the private values for one participant usually differ from the private values of the other participants. Your private values are the same for every round in the experiment.

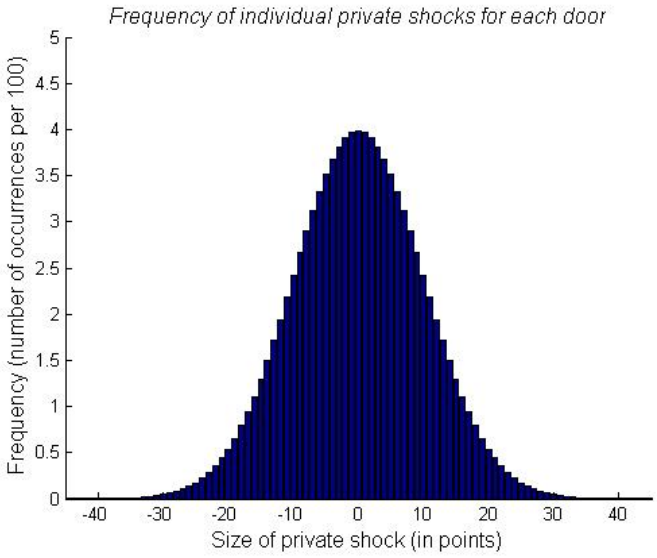


Figure 12

Social value

Your social value in a round depends on how many other people in your group make the same door choice as you. You gain if the majority of the other participants make the same choice as you, but you make a loss if the majority makes the other choice. Specifically, you **gain 8 points** for every person who makes the **same** choice as you, but you **lose 8 points** for every person who makes the **opposite** choice to you. As there are five other people in your group, you can get a maximum social value of 40 points if everyone chooses the same door as you, or you can maximally lose 40 points if everyone chooses the other door to you.

The other participants in your group face the same decision as you do. That is, they receive similar information as you do (although their private values will most likely differ), they also choose between Door A and Door B and they make money in the same way as you do.

Example

In this game, there are 5 other participants in your group. So, for example, if you choose Door A with a common value of 80 points, a private value of -10 points and 4 others also choose Door A, your payoff equals the **common value** plus your **private value** (80 ? 10 = 70) plus a **social value** (32 - 8 = 24), for a **total of 94 points**.

If on the other hand you choose Door B with a common value of 40 points and a private value of 20 points, and 5 others also choose Door B, your payoff equals the **common value** plus your **private value** (40 + 20 = 60) plus a **social value** of 40 points, for a **total of 100 points**.

Sequence of events

Summing up, each round is characterised by this sequence of events:

- At the start of each round, you are told your constant private values for the doors.
- At the start of each round, you are told the new common values for the doors.
- You make your choice between Door A and Door B.
- At the end of a round, you are told the number of your group members who made each choice, what the social values were for those who chose each door, and you are informed of your payoff in that round. Each round's payoff is the sum of your chosen door's **common value**, your **private value** and your chosen door's **social value**.

Other participants face exactly the same sequence of events.

You can always see the history of the groups choices for all rounds up to that point at the bottom of your screen. You can also always see the sum of the number of points that you earned so far at the top left corner of your screen.

On the next screen you will be requested to answer some control questions. Please answer these questions now.

8 Appendix C: Table of results

Table 4: Key performance indicators by group

Group number	N	J	ρ_{50}	$\bar{\rho}_{(45-50)}$	$\bar{\rho}_{all}$	$\bar{\rho}_{(t\geq 26)}$	t_{switch}	Earnings(€)	Leaders	Testers
1	6	4	.00	.00	.04	.03	-	8.76	0.00	100.00
2	6	4	.17	.06	.03	.03	-	8.76	0.00	66.67
3	6	4	.00	.00	.03	.04	-	8.84	0.00	83.33
4	6	4	1.00	.97	.27	.49	38	8.39	33.33	66.67
5	6	4	1.00	1.00	.42	.79	30	8.91	16.67	83.33
6	6	4	1.00	1.00	.46	.85	28	8.69	33.33	66.67
7	6	4	1.00	1.00	.32	.62	35	9.21	33.33	33.33
8	6	4	1.00	.94	.48	.84	17	8.09	16.67	66.67
9	6	8	.00	.00	.03	.02	-	12.39	16.67	66.67
10	6	8	.00	.00	.05	.07	-	11.77	16.67	50.00
11	6	8	.00	.00	.02	.01	-	12.78	0.00	50.00
12	6	8	.00	.00	.05	.06	-	11.77	33.33	16.67
13	6	8	.00	.00	.01	.00	-	12.90	0.00	33.33
14	6	8	.00	.00	.02	.00	-	12.49	16.67	50.00
15	6	8	.00	.00	.03	.03	-	12.38	16.67	33.33
16	6	8	.00	.00	.02	.01	-	12.72	0.00	33.33
17	11	4	.09	.02	.01	.01	-	9.07	0.00	54.55
18	11	4	1.00	.98	.40	.76	32	8.97	45.45	54.55
19	11	4	.91	.41	.10	.18	49	8.45	45.45	45.45
20	11	4	1.00	.98	.32	.62	36	8.99	45.45	45.45
21	11	4	.00	.00	.02	.01	-	9.00	0.00	63.64
22	11	4	.09	.05	.06	.09	-	8.56	36.36	36.36
23	11	4	.18	.09	.07	.12	-	8.48	27.27	63.64
24	11	8	.00	.02	.02	.02	-	12.82	9.09	36.36
25	11	8	.00	.00	.03	.04	-	12.28	9.09	36.36
26	11	8	.09	.06	.02	.02	-	12.80	9.09	36.36
27	11	8	.00	.02	.03	.03	-	12.38	18.18	18.18
28	11	8	.00	.00	.04	.04	-	12.04	27.27	27.27
29	11	8	.00	.00	.01	.00	-	13.09	0.00	27.27
30	11	8	.09	.02	.02	.02	-	12.58	0.00	63.64

Notes: Values are averages group values. Earnings do not include the €3 show-up fee. ρ_{50} = final group proportion choosing Door A. $\bar{\rho}_{(45-50)}$ = average ρ across the last final six rounds. $\bar{\rho}_{all}$ = average ρ across all rounds. $\bar{\rho}_{(t\geq 26)}$ = average ρ from round 26, when the common value of Door A becomes larger than that of Door B. t_{switch} considers only those groups that switched to Door B by round 50. *Testers* and *Leaders* are percentages of the respective individual types: Testers deviate from the group norm in one round before reverting back to the group choice, while Leaders deviate from the group norm in at least two consecutive rounds. Highlighted rows are those groups defined as having switched to Door A by the end of the experiment.

References

Abbink, Klaus, Brandts, Jordi, Herrmann, Benedikt, & Orzen, Henrik. 2010. Intergroup Conflict and Intra-group Punishment in an Experimental Contest Game. *American Economic Review*, **100**(1), 420–47.

Akerlof, George A., & Kranton, Rachel E. 2000. Economics and Identity. *The Quarterly Journal of Economics*, **115**(3), 715–753.

Arrow, Kenneth. 1970. Political and Economic Evaluation of Social Effects and Externalities. *Pages 1–30 of: The Analysis of Public Output*. NBER Chapters. National Bureau of Economic Research, Inc.

Asch, S. E. 1956. Studies of independence and conformity: A minority of one against a unanimous majority. *Psychological Monographs*, **70**.

Brock, William A., & Durlauf, Steven N. 2001. Discrete Choice with Social Interactions. *The Review of Economic Studies*, **68**(2), pp. 235–260.

Cachon, Gerard P., & Camerer, Colin F. 1996. Loss-Avoidance and Forward Induction in Experimental Coordination Games. *The Quarterly Journal of Economics*, **111**(1), 165–94.

Charness, Gary, Rigotti, Luca, & Rustichini, Aldo. 2007. Individual Behavior and Group Membership. *American Economic Review*, **97**(4), 1340–1352.

Chaudhuri, Ananish, Schotter, Andrew, & Sopher, Barry. 2009. Talking Ourselves to Efficiency: Coordination in Inter-Generational Minimum Effort Games with Private, Almost Common and Common Knowledge of Advice. *Economic Journal*, **119**(534), 91–122.

D'Cruz, Andrea. 2005. The Handshake. *Assyria Times*, **Accessed 20 June 2013**(May 5).

Eggertsson, Thrainn. 2001. Norms in Economics, With Special Reference to Economic Development. *Chap. 3, pages 76–104 of:* Hechter, M., & Opp, K.D. (eds), *Social Norms*. Russell Sage Foundation.

Fershtman, Chaim, & Gneezy, Uri. 2001. Discrimination in a Segmented Society: An Experimental Approach. *The Quarterly Journal of Economics*, **116**(1), 351–377.

Gage, Anastasia J. 1998. Sexual Activity and Contraceptive Use: The Components of the Decisionmaking Process. *Studies in Family Planning*, **29**(2), pp. 154–166.

Gneezy, Uri, Leonard, Kenneth, & List, John. 2009. Gender Differences in Competition: Evidence From a Matrilineal and a Patriarchal Society. *Econometrica*, **77**(5), 1637–1664.

Hargreaves Heap, Shaun, & Zizzo, Daniel. 2009. The Value of Groups. *American Economic Review*, **99**(1), 295–323.

Hart, P. Sol, & Nisbet, Erik C. 2011. Boomerang Effects in Science Communication: How Motivated Reasoning and Identity Cues Amplify Opinion Polarization About Climate Mitigation Policies. *Communication Research*.

Hechter, M., & Opp, K.D. 2001. *Social Norms*. Russell Sage Foundation.

Inzlicht, Michael, & Kang, Sonia K. 2010. Stereotype threat spillover: how coping with threats to social identity affects aggression, eating, decision making, and attention. *Journal of personality and social psychology*, **99**(3), 467–81.

Kinzig, Ann P., Ehrlich, Paul R., Alston, Lee J., Arrow, Kenneth, Barrett, Scott, Buchman, Timothy G., Daily, Gretchen C., Levin, Bruce, Levin, Simon, Oppenheimer, Michael, Ostrom, Elinor, & Saari, Donald. 2013. Social Norms and Global Environmental Challenges: The Complex Interaction of Behaviors, Values, and Policy. *BioScience*, **63**(3), 164–175.

Knez, Marc, & Camerer, Colin. 1994. Creating Expectational Assets in the Laboratory: Coordination in Weakest?Link Games. *Strategic Management Journal*, **15**, 101–119.

Kopanyi-Peuker, Anita, Offerman, Theo, & Sloof, Randolph. 2015. Probation or Promotion? The Fear of Exclusion Improves Team-Production. *Discussion Paper, University of Amsterdam*.

Lessig, L. 1995. *The Regulation of Social Meaning*. The University of Chicago.

Luhan, Wolfgang J, Kocher, Martin G, & Sutter, Matthias. 2009. Group polarization in the team dictator game reconsidered. *Experimental Economics*, **12**(1), 26–41.

Mackie, Gerry. 1996. Ending Footbinding and Infibulation: A Convention Account. *American Sociological Review*, **61**(6), pp. 999–1017.

Mannix, Elizabeth A., & Loewenstein, George F. 1994. The Effects of Interfirm Mobility and Individual versus Group Decision Making on Managerial Time Horizons. *Organizational Behavior and Human Decision Processes*, **59**(3), 371–390.

Milgram, S. 1974. *Obedience to authority: an experimental view*. Harper & Row.

Munshi, Kaivan, & Myaux, Jacques. 2006. Social norms and the fertility transition. *Journal of Development Economics*, **80**(1), 1–38.

Neyfakh, Leon. 2013. Do handshakes make you sick? *The Boston Globe*, **Accessed 20 June 2013**(February 17).

Postlewaite, Andrew. 2010 (May). *Social Norms and Preferences, Chapter for the Handbook for Social Economics, Edited by J. Benhabib, A. Bisin and M. Jackson*. PIER Working Paper Archive 10-031. Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

Prentice, Deborah A. 2012. The Psychology of Social Norms and the Promotion of Human Rights. *Pages 23–46 of:* Goodman, R., Jinks, D., & Woods, A.K. (eds), *Understanding Social Action, Promoting Human Rights.* Oxford University Press.

Sandstrom, Marlene, Makover, Heather, & Bartini, Maria. 2013. Social context of bullying: Do misperceptions of group norms influence children's responses to witnessed episodes? *Social Influence*, **8**(2-3), 196–215.

Schroeder, Christine M., & Prentice, Deborah A. 1998. Exposing Pluralistic Ignorance to Reduce Alcohol Use Among College Students1. *Journal of Applied Social Psychology*, **28**(23), 2150–2180.

Sherif, Muzafer. 1936. *The Psychology of Social Norms.* Harper and Brothers.

Singh, Ajit, & Dhumale, R. 2000. *Globalization, Technology, and Income Inequality: A Critical Analysis.* Research Paper. World Institute for Development Economics Research.

Tajfel, H., & Turner, J. C. 1979. An integrative theory of intergroup conflict. *Pages 33–47 of:* Austin, W. G., & Worchel, S. (eds), *The social psychology of intergroup relations.* Brooks/Cole.

Tajfel, H., & Turner, J. C. 1986. The social identity theory of intergroup behaviour. *Pages 7–24 of:* Austin, W. G., & Worchel, S. (eds), *Psychology of Intergroup Relations.* Nelson-Hall.

Van Huyck, John B, Battalio, Raymond C, & Beil, Richard O. 1990. Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure. *American Economic Review*, **80**(1), 234–48.

Weber, Roberto A. 2006. Managing Growth to Achieve Efficient Coordination in Large Groups. *American Economic Review*, **96**(1), 114–126.

Wenzel, Michael. 2005. Misperceptions of social norms about tax compliance: From theory to intervention. *Journal of Economic Psychology*, **26**(6), 862–883.

Yan Chen, Sherry Xin Li. 2009. Group Identity and Social Preferences. *The American Economic Review*, **99**(1), 431–457.

Young, H. Peyton. 2008. Social Norms. *In:* Durlauf, Steven N., & Blume, Lawrence E. (eds), *The New Palgrave Dictionary of Economics.* Basingstoke: Palgrave Macmillan.

Young, S.D., Hlavka, Z., Modiba, P., Gray, G., Van Rooyen, H., Richter, L., Szekeres, G., & Coates, T. 2010. HIV-related stigma, social norms, and HIV testing in Soweto and Vulindlela, South Africa. *Journal of Acquired Immune Deficiency Syndromes*, **55**(5), 620–624.

Zimbardo, P.G. 1972. *The Psychology of Imprisonment: Privation, Power and Pathology.* Stanford University.