Social Change and the Conformity Trap*

James Andreoni† Nikos Nikiforakis‡ Simon Siegenthaler§

December 31, 2017

Preliminary Draft

Abstract

The ability of societies to adapt to a changing world is critical for welfare. We present evidence from a new laboratory experiment in which individuals' preferences change over time and there is a pressure to conform to the behavior of the majority. In the baseline environment, groups fail to adapt behavior to the changing circumstances despite the common knowledge that it has become inefficient and undesirable to almost all group members. We then explore factors that may help avoid these *conformity traps*, such as organizing opinion polls, expediting social feedback, increasing tolerance, and rewarding instigators of change. We also show that change is hindered by risk aversion but promoted by individuals who derive utility from defying the pressure to conform.

JEL Classification: C92, D60, D70

Keywords: Conformity, Coordination Failure, Equilibrium Selection, History Dependence, Social Norms

*We would like to thank Vincent Leah-Martin and Seung-Keun Martinez for excellent research assistance. We thank Olivier Bochet, Aurélie Dariel, Nisvan Erkal, Guillaume Fréchette, Marie Claire Villeval, Thomas Palfrey, Karl Schlag, Eric Snowberg and Tom Wilkening for helpful comments. We also thank participants of the 2015 Workshop on Behavioral Political Economy at NYU Abu Dhabi, the 2016 NYU Global Network Experimental Social Sciences Workshop, the 2016 ESA World Meeting in Jerusalem, the 2016 ESA European Meeting in Bergen, the 2017 European Workshop in Experimental and Behavioral Economics in Bologna, and seminar participants at the Wissenschaftszentrum Berlin für Sozialforschung, the University of Texas at Dallas, and the universities of Gothenburg, Melbourne, Ottawa and Vienna. All aspects of the experiment were reviewed by the Internal Review Board at New York University Abu Dhabi (#027-2015) and at the University of California, San Diego (#150689).

†Department of Economics, University of California, San Diego, La Jolla, CA 92093, USA. E-mail: andreoni@ucsd.edu. Phone: +1 858-534-3832.

‡Division of Social Science, New York University Abu Dhabi, PO Box 129188, United Arab Emirates. E-mail: nikos.nikiforakis@nyu.edu. Phone: +971 26285436.

§Naveen Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080, USA. E-mail: simon.siegenthaler@utdallas.edu.

"The reasonable man adapts himself to the world; the unreasonable one persists in trying to adapt the world to himself. Therefore all progress depends on the unreasonable man."

— George Bernard Shaw (1903)

1 Introduction

A defining feature of the last 100 years has been the continuous advances in technology, transportation, and telecommunications. These developments have eroded traditional boundaries of human interaction, bringing people from different cultures closer than ever, challenging in the process long-established social values, norms and institutions. Against this background, the ability of societies to rapidly adapt, taking advantage of new opportunities and avoiding potential threats, is critical for the welfare of their members. A fundamental question regarding social change is whether it occurs when it is socially beneficial or whether it is slow to ignite (e.g., North, 1990; Ostrom, 2000; Acemoglu et al., 2005; Weinstein, 2010).

In most instances, assessing the efficiency of social change is problematic due to the paucity of reliable data on individual preferences in daily life and the difficulty of aggregating them. Yet, in some cases, it is easy to see that socially beneficial change can take too long to materialize. Female genital mutilation and child marriage are two prominent examples. Despite the fact that these practices are illegal, they are supported by strong social norms in some countries and thus persist to this day (e.g., Mackie, 1996; Bicchieri, 2006; Bicchieri and Muldoon, 2011). Other examples include bans on interracial marriage, norms of personal revenge (e.g., dueling), or the custom of foot binding (e.g., Elster, 1989; Coleman, 1994). A recent, striking example is offered by the case of Harvey Weinstein, which we will discuss below.

In this paper, we ask what can cause entrenchment of unproductive or inefficient paradigms, like in the examples above, and what forces can spark and accelerate widely beneficial change. We study these social shifts in an unlikely arena—a laboratory environment where 20 human subjects gather for 90 minutes to play coordination games under slightly different rules. We use one such session as a single observation. Despite being small in scale and artificial in nature, we nonetheless find these exercises to be intriguing and illuminating on the question of what brings social change and what impedes it. They also allow us to establish a set of clearly constructed examples of interventions in society that can help expedite change.

Our first task is to find an environment where the inherited paradigm is nearly impossible to escape despite the common knowledge that it has become inefficient and undesirable to almost the whole of society. We use the following game. There is a set of players which belongs to a large group. In every round, each player is randomly matched with another and must choose between two colors: *blue* or *green*. If the players fail to coordinate on the same color, they suffer a "disunity penalty." This penalty is increasing in the number of people in the group choosing the *opposite* color such that everyone choosing *blue* and everyone choosing *green* are both equilibria of the one-period game.[1]

[1]The need to coordinate is a key feature in many social interactions whether they involve social norms, customs or

Preferences evolve over time reflecting the arrival of new information. At the start of the experiment, all group members prefer *blue*, but preferences change gradually such that, at some point, virtually all group members would prefer switching to *green*. All aspects of the game are common knowledge. Key to our game is that "pioneers of social change" incur disproportionately large costs for deviating from the status quo, creating incentives to wait for others to deviate first. If everyone acts in this way, however, societies get caught in a *conformity trap*, where socially beneficial adaptation does not occur.

Many elements of our lab study were in full display in the recent mass social change that has become known as the *#MeToo* movement. First, a beloved actor and comedian Bill Cosby was convicted in federal court of sexually preying on women. Then Jimmy Saville, a British personality well known as an entertainer for children and a supporter of charities that advocate for children, was exposed posthumously for hundreds of cases of child sexual abuse. Then in the fall of 2017 the cascade began with the case of Harvey Weinstein. For years it had been an open secret that Harvey Weinstein, an influential producer in Hollywood, had abused his power as a gatekeeper to cinematic success by behaving in sexually abusive and sometimes violent ways with actresses, often threatening the careers of those who might expose him.[2] However, outside of Hollywood, few people were aware of the crimes implicitly endorsed by consuming his films. Then, on October 5, the *New York Times* used both legal documents and personal interviews with now famous actors to chronicle the length and severity of the abuse by Weinstein. Immediately, others not discussed in the article spilled forward to describe their own stories. Two days later Weinstein resigned his position in Mirimax. As of mid-December 2017, 84 women have come forward to tell their stories.[3] Yet, for a long time no one was willing to come forward, or those that did were victimized again through professional retribution.

Our game can be applied to these cases of sexual abuse. The players are the set of people who have knowledge of abusive behavior, whether victims themselves or not. Each of them must propose whether to keep quiet (*blue*) or to speak out (*green*). Preferences change at a (commonly-known) rate which reflects the arrival of new information (e.g., Harvey Weinstein assaulted person X) or evolving beliefs that speaking out can bring positive change to the way things are done. At regular intervals the players meet up (which we model as randomly in pairs) and discuss whether in their opinion they should keep quiet or speak out. If they agree, they go on with their day, happy that they met a like-minded person. If they disagree, they experience a psychic cost which relates to their fear of either being exposed as one willing to talk, or by association with one who is. The way events unfolded in the Weinstein case suggests that the cost of speaking out are largest for individuals who do so first, and decreases in the number of others who have come forward.

The experimental results reveal that social change in our lab environment is slow to ignite and often fails to occur altogether, even when it is common knowledge that everyone would benefit from it. Paradoxically, while social pressure is sometimes necessary for the enforcement of efficiency-enhancing norms—for instance when sanctioning free-riders in public good environments (e.g., Fehr and Gächter,

conventions and miscoordination entails individual costs, e.g., due to social sanctions or feelings of guilt or not belonging.

[2]See the article by CNN, "Mirimax Insider: Everybody Knew about Weinstein's Behavior," http://money.cnn.com/2017/10/17/media/scott-rosenberg-harvey-weinstein/index.html.

[3]See "A comprehensive timeline of the Harvey Weinstein allegations," Vogue Magazine, December 7, 2017.

2000)—it can also limit a group's ability to adapt to changing circumstances. This implies that central interventions may be necessary for societies to advance. To this end, in further treatments, we explore factors that can help groups escape the conformity trap. Our results show that promoting tolerance, organizing opinion polls, and rewarding pioneers of change are interventions that can increase the likelihood of change, but none of them is a panacea. We are also interested in the role of information. Can the faster spread of information, such as through social media, help spark change? Can an objective and public recording of opinions, such as through a credible public opinion survey or a surprise election result, be the catalyst to precipitate change?

The paper proceeds as follows. In the next section, we provide a brief discussion of the related literature. Section 3 introduces the social change game. Section 4 presents the experimental design. Section 5 presents an analytical framework and a set of hypotheses. Section 6 discusses our main results, explores implications for welfare, and examines some characteristics of instigators of change. Finally, Section 7 concludes.

2 Related Literature on Conformity and Social Norms

Our findings show that the pressure to conform can prevent societies from adapting in a changing world. In this regard, our study is related to theoretical investigations that model norms as equilibria in coordination games (e.g., Brock and Durlauf, 2001). Inefficient norms may emerge when the efficient norm is associated with greater risk in case of miscoordination than the inefficient one, i.e., when there is a tension between payoff and risk dominance (e.g., Kandori et al., 1993; Young, 1993); when members of overlapping generations must coordinate with each other (Acemoglu and Jackson, 2015); or when they have heterogeneous tastes (Michaeli and Spiro, 2017).[4]

Our study is also related to the literature investigating institutional change more broadly (e.g., North, 1990; Williamson, 2000; Greif and Laitin, 2004). Acemoglu and Robinson (2008) emphasizes the role of powerful elites in preventing change. Arthur (1989) lists different reasons for the persistence of inefficient institutions: incomplete and imperfect information, adaptive expectations where increased prevalence of an institution enhances beliefs of future prevalence, and network effects where the value of an institution depends on the number of other people participating in it. Network externalities have also been used to explain technological lock-ins in markets (David, 1985; Farrell and Saloner, 1985; Katz and Shapiro, 1985), although their empirical relevance is disputed (Liebowitz and Margolis, 1994, 1995). Our set-up also exhibits a form of network externalities which has to do with the disunity penalty falling in the number of non-conformists.

Other authors have relied on laboratory experiments to study the persistence of inefficient social institution. Wilkening (2016) shows that institutions that emerge to alleviate moral hazard such as costly certification may persist after they cease to be efficient due to a problem of incomplete information.

[4] In a laboratory experiment, Friedman (1996) finds that risk dominance fares worse than predicted in Kandori et al. (1993) and Young (1993) as an equilibrium selection criterion as players' deviations are often not random trembles, but rather deliberate attempts to persuade other players to seek Pareto superior outcomes.

In the context of technology adoption, Hossain and Morgan (2009) and Hossain et al. (2011) find that in a setting with network externalities and complete information, groups manage to coordinate on the Pareto superior technological platform, even if initially forced to choose inferior platforms. That is, they observe no lock-ins. This finding is replicated in Heggedal and Helland (2014).[5] The study most closely related to ours is Smerdon et al. (2016). The studies were independently conceived and conducted. Both conceptualize norms as equilibria in a coordination game and investigate situations in which preferences over two alternatives change over time.[6] However, the mechanism leading to the persistence of inefficient social norms in Smerdon et al. (2016) differs from ours. In their study participants don't know whether or in what way preferences change. As a result, many subjects seem to erroneously believe that the majority of other group members prefers the status quo. This is often referred to as pluralistic ignorance. In contrast, in our experiment, subjects' preferences change at a steady, commonly-known rate; norm adaptation is hampered by the higher costs for instigators of change.

3 The Social Change Game

Our game models a situation in which individuals' preferences change over time and interactions are governed by a social norm. There is a group of n players. Preferences are given by a player's type $\theta = \{B, G\}$. Type B players prefer *blue* (b) and type G players prefer *green* (g). Denote the value of a type θ player for choosing $c = \{b, g\}$ by $v_\theta(c)$. We have $v_B \equiv v_B(b) - v_B(g) > 0$ and $v_G \equiv v_G(g) - v_G(b) > 0$. Players interact in periods $t = 1, \ldots, T$. In each period, they are matched into pairs at random and choose action *blue* or *green*.

In line with the existence of norm, players receive a penalty if they choose a different color than most other players in the group. Specifically, if two players in a match miscoordinate—that is, they choose different colors—both players receive a *disunity penalty*. The penalty is given by the number of players in the group who choose the opposite color multiplied by a parameter p. Thus, in a given period, the payoff $\Pi_{\theta_i}^{c_i c_j}$ of player i when choosing c_i and being matched with player j choosing c_j is:

$$\begin{aligned}
\Pi_{\theta_i}^{bb} &= v_{\theta_i}(b) \\
\Pi_{\theta_i}^{gg} &= v_{\theta_i}(g) \\
\Pi_{\theta_i}^{bg}(n_g) &= v_{\theta_i}(b) - n_g p \\
\Pi_{\theta_i}^{gb}(n_b) &= v_{\theta_i}(g) - n_b p
\end{aligned}$$

where n_g is the total number of other players choosing *green* and $n_b = n - 1 - n_g$. Throughout the

[5] In a variant of the minimum-effort game (Van Huyck et al., 1990), in which the tension between payoff and risk dominance often leads to convergence to the inefficient equilibrium, Brandts and Cooper (2006) show that groups can overcome coordination failure if the benefits for coordinating on high effort are simultaneously increased for all group members. See also Gërxhani and Bruggeman (2015) and Masiliūnas (2017) for experiments examining the ability of social groups to overcome coordination failures.

[6] Apart from these commonalities, Smerdon et al. (2016)'s experimental setting differs from ours, particularly in terms of the matching technology, the way preferences change, and the information structure.

Table 1: Expected Fraction of Subjects Preferring *Green* (Type G)

Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Type G	0%	10%	19%	27%	34%	41%	47%	52%	57%	61%	65%	69%	72%	75%	77%	79%
Period	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	
Type G	81%	83%	85%	86%	88%	89%	90%	91%	92%	93%	94%	94%	95%	95%	96%	

paper we focus on situations in which $(n-1)p > v_B$ and $(n-1)p > v_G$. That is, choosing the preferred color is never a dominant strategy.

A key feature of the social change game is that players' preferences over colors change over time. The game starts in period 1 when all players are type B. From period 2 onwards, in each period up to and including period T, each player who is still type B switches to type G with probability r. The rate of change r is *common knowledge*. Players who have become type G never switch back to type B. The expected fraction of type G players in period t is thus $f_{G,t} = 1 - (1-r)^{t-1}$. In the experiment, the number of periods is $T = 31$ and the rate of change is $r = 0.1$. Table 1 shows the expected fraction of type G subjects in each period. By period 8, a majority of individuals in the group prefer *green* (in expectation), and the support for *green* grows to 81% in period 17 and exceeds 90% after period 23.

Hence, in period 1, players face a coordination game with fully aligned interests; it is commonly known that everyone is type B. Players therefore likely coordinate on the payoff-dominant choice *blue* (Harsanyi and Selten, 1988). As time proceeds, the probability that subjects are type G increases, such that after an intermediate phase during which both types are common, players become increasingly certain that virtually everyone prefers *green*. The situation is then as in period 1, except that now *green* is payoff-dominant (and there is a history of previous interactions).[7]

4 The Experiment

The experimental design, summarized in Table 2, follows the social change game introduced above.

4.1 Experimental Treatments

Baseline: The Baseline treatment implements the social change game with parameters as given in Figure 1. That is, in each session subjects are randomly matched into pairs in $T = 31$ periods and choose between *blue* (b) and *green* (g). Initially, everyone prefers *blue* but there is a rate of change $r = 10\%$ such that more and more subjects prefer *green*. Choosing the preferred color yields a payoff of 30 and choosing the other color a payoff of 20. In case of miscoordination, subjects incur a disunity

[7] There is no generally agreed-on definition of risk dominance that applies to our setting. See Peski (2010) for a generalization of the concept of risk dominance where, when applied to our setting, *green* is the risk dominant equilibrium in period 1 and 31.

Table 2: Experimental Design

Treatment	Subjects (Sessions)	Difference to Baseline
Baseline	120 (6)	Social change game with parameters as in Figure 1.
High Return	120 (6)	Higher payoff for type G if choosing g: $v_G(g) = 50$.
Low Penalty	120 (6)	Lower disunity penalty parameter: $p = 1$.
Endogenous Penalty	120 (6)	Subjects choose disunity penalty: $p = \{1, 4, 7\}$.
Reward	120 (6)	Initiators of change receive highest earnings.
Poll	120 (6)	Poll about preferred color in period 14.
Fast Information	120 (6)	Immediate feedback about others' color choices.
Small Group	60 (6)	Smaller group: $n = 10$. Same expected penalty: $p = 8.44$.

Sessions were run at the University of California, San Diego, in summer 2015 (treatment Reward in fall 2016). Total number of participants: 900. Average payment: $36.4 (US).

Figure 1: Social Change Game with Baseline Parameters

Both players are type B

	b	g
b	30 30	$30 - n_g * 4$ $20 - n_b * 4$
g	$20 - n_b * 4$ $30 - n_g * 4$	20 20

$\xrightarrow{\text{r=10\%}}$

Row: type G, Col: type B

	b	g
b	20 30	$20 - n_g * 4$ $20 - n_b * 4$
g	$30 - n_b * 4$ $30 - n_g * 4$	30 20

$\xrightarrow{\text{r=10\%}}$

Both players are type G

	b	g
b	20 20	$20 - n_g * 4$ $30 - n_b * 4$
g	$30 - n_b * 4$ $20 - n_g * 4$	30 30

Groups consist of $n = 20$ subjects randomly matched into pairs over $T = 31$ periods. The rate of change is $r = 10\%$. The first (top) entry in each cell is the row player's payoff; the second (bottom) entry the column player's payoff. The variable n_c is the total number of players choosing $c = \{b, g\}$.

penalty of $p = 4$ for each subject in the group choosing the other color. At the end of a period, subjects are informed about the action chosen by their matched subject (but not the other's type). Players are also informed about their earnings and the total number of players in the group who chose *blue* and *green*, but this feedback is given with a delay of one period.[8]

High Return: In High Return, we vary one of the key environmental parameters: the returns to social change. The payoff of type G subjects if choosing *green* is increased from 30 to 50. Type B subjects still earn 30 for choosing *blue* and the payoff for choosing the color not in line with one's type remains at 20.

Low Penalty: Treatment Low Penalty is identical to Baseline, except that the disunity penalty parameter is lowered to $p = 1$ ($p = 4$ in Baseline). This alleviates conformity effects. We interpret the lower penalty in terms of a more tolerant society.

Endogenous Penalty: In each period of the Endogenous Penalty treatment, besides choosing a color, subjects also choose a value of the disunity penalty to be paid by their matched participant in

[8]Delaying feedback about aggregate group behavior is consistent with our modeling choice of pairwise interactions. Information about the choice of the other person in a match is instantly revealed, but behavior in the rest of the society is disseminated only over time. The speed of feedback is a potentially important variable affecting the probability of conformity traps to occur. In treatment Fast Information, we will remove the delay in feedback.

the event of miscoordination. The available values are $p = \{1, 4, 7\}$. This introduces the possibility for type B subjects to encourage non-conformity (other subjects choosing *green*) by keeping penalties low, even if they are themselves still choosing *blue*.

Reward: Let the majority color be the one chosen by 11 or more subjects in the final period. A reward is received by the four subjects who have persisted the longest in choosing the majority color, irrespective of whether the majority color is *blue* or *green*.[9] Initiating change thus promises a reward, but it is also risky. Outside the laboratory, rewards received by pioneers of change may take the form of privileged social positions, political power, fame, prizes, decorations, and monuments.

To trigger a change from *blue* to *green*, type G players need to unite when they are sufficiently numerous. Empirically, however, coordinating on a specific period of change may be difficult. The following three treatments explore environments in which this coordination problem is alleviated.

Poll: A poll is conducted in period 14. Subjects are asked what color they would prefer people in their matching group chose in the next rounds. After learning how many people answered *blue* and how many *green*, all subjects make their actual color choices for period 14. All aspects of the poll are explained in the experiment instructions and subjects are aware in period 1 that there will be a poll. In period 14, in expectation 75% of the subjects prefer *green* (type G), and the probability that the group will have a majority prefering *green* is 98.5%.

Fast Information: In treatment Fast Information, subjects receive immediate feedback about the number of people who chose *blue* and *green* at the end of each period. Compared to Baseline, where this information is delayed by one period, treatment Fast Information should improve subjects' ability to signal a willingness to lead the change. Outside the laboratory, faster information dissemination may be a consequence of improved communication channels such as television and social media.

Small Group: The group size is reduced to $n = 10$ ($n = 20$ in Baseline). In smaller groups strategic uncertainty may be alleviated and coordination facilitated. Importantly, we simultaneously increase the disunity penalty parameter to $p = 8.44$ to keep the expected cost of miscoordination the same as in Baseline. For instance, the disunity penalty of an isolated deviation to *green* when everyone else is choosing *blue* is $19 * 4 = 76$ in Baseline and $9 * 8.44 = 76$ in Small Group.

4.2 Eliciting Individual Characteristics

At the end of a session, we elicited subjects' risk aversion using a task taken from Andreoni and Harbaugh (2016).[10] We also elicited preferences for non-conformity by asking subjects to rate statements taken from a scale measuring psychological reactance developed by Hong and Faedda (1996).[11]

[9] If there is no majority color in period 31 (each color is chosen by 10 subjects), no rewards are distributed.

[10] Subjects had to pick one of six lotteries: (a) 8 in 10 chance to win \$2 (US), (b) 7 in 10 chance to win \$3, (c) 6 in 10 chance to win \$4, (d) 5 in 10 chance to win \$5, (e) 4 in 10 chance to win \$6, and (f) 3 in 10 chance to win \$7. Options (a) to (f) order subjects by risk aversion, with (a) revealing the greatest risk aversion, (d) revealing risk neutrality (it maximizes expected value), and (f) is the most risk loving choice.

[11] The statements are: "I become angry when my freedom of choice is restricted," "It disappoints me to see others submitting to standards and rules," "When someone forces me to do something, I feel like doing the opposite," "I become frustrated when I am unable to make free and independent decisions," "I find contradicting others stimulating,"

4.3 Procedures

The experiment was conducted at the Economics Laboratory of the University of California, San Diego (UCSD) using z-Tree (Fischbacher, 2007). A total of 900 subjects participated in the experiment. We ran six sessions for each treatment. Each subject participated only in one session. All sessions consisted of 20 participants, except the Small Group treatment, which had 10 individuals per session. The sessions were run between September 2015 and October 2016. Participants were students at UCSD from various disciplines. The mean age was 20 years and 53% of the participants are female.

Written instructions were distributed at the beginning of the experiment; the experimenter also read them aloud. The instructions for all treatments can be found on the authors' web sites and in the online appendix. The experiment started after all participants had correctly answered a set of control questions included in the instructions. Sessions lasted less than 70 minutes. Earnings were given in experimental currency units (ECU) and converted into US Dollars at the end of the experiment (1 ECU = \$0.03). Subjects were paid the sum of their earnings over all 31 periods. Payments averaged \$36.4 per subject, including a show up fee of \$10.

5 Analytical Framework and Hypotheses

The social change game has a large number of Perfect Bayesian equilibria.[12] This section proposes a natural way of selecting a specific equilibrium, motivating the different experimental treatments and allowing us to derive behavioral hypotheses. A full dynamic model of the process of social adaptation is beyond the scope of this study, although the experimental results should be informative in developing an (economic) theory of social change.

5.1 Efficient Period of Change

For sufficiently large disunity penalties, the behavior that maximizes the sum of ex-ante expected payoffs is when all players switch to *green* in the same period in order to avoid miscoordination. To be more precise, the efficient switching period is given by

$$\min\left(t : f_{G,t} \geq v_B/(v_B + v_G)\right) \tag{1}$$

where $f_{G,t}$ is the expected fraction of type G players in period t. For the parameters in the baseline condition of the experiment, expression (1) implies that all players should switch to choosing *green* in

"Regulations trigger a sense of resistance in me," "The thought of being dependent on others aggravates me," "It irritates me when someone points out things which are obvious to me," "I am content only when I am acting of my own free will" and "I resist the attempts of others to influence me." See Goldsmith et al. (2005) for a detailed discussion of the scale in the context of conformity.

[12] The set of Perfect Bayesian equilibria set contains (but is not limited to) any combination of one-period Bayes Nash equilibria (BNE). In any period, everyone choosing *green* or everyone choosing *blue* are always BNE. There can also be other BNE in which either a subset of type G (type B) players choose *green* (*blue*), while everyone else chooses *blue* (*green*).

period 8 out of 31, as soon as a majority of the subject has become type G.[13]

5.2 Risk Aversion

We allow for risk aversion. Each player i draws a risk parameter x_i from the normal distribution $\mathcal{N}(\mu, \alpha^2)$. The corresponding utilities in a given period are denoted by $u_i(\Pi^{c_i c_j}_{\theta_i})$. Switching to *green* is risky, as players are randomly matched and, despite the commonly known rate of change r, players don't know the exact number of type G players in each period. Risk aversion thus tends to discourage deviations from the *blue* norm.

5.3 Equilibrium Period of Change

Our key question is as follows: if we assume that all players choose *blue* in period 1 (when everyone is type B), when will the growing number of type G players be sufficiently large to trigger a change in group behavior to the *green* norm? An intuitive approach to the problem is to say that social change occurs as soon as there is a group of type G players who prefer jointly deviating to *green* (assuming the remaining players choose *blue*) to the alternative in which everyone keeps choosing *blue*.

Notice that type B players will never be part of the group challenging the *blue* norm. This implies that highly risk averse type G players whose main concern is to avoid miscoordination also prefer to delay social change. The group of players initiating change must hence consist of type G individuals with low levels of risk aversion. More formally, we are looking for the first period in which there exists a threshold risk level $\hat{x}$ such that all type G players with $x_i \leq \hat{x}$ are willing to choose *green*, even if the remaining players choose *blue*. It suffices to check if the marginal player m with $x_m = \hat{x}$ prefers to choose *green*, which yields the following condition:

$$\sum_{n_G=0}^{n-1} p(n_G, t) \sum_{n_g=0}^{n_G} q_m(n_g, n_G) \left(\frac{n_g}{n-1} u_m(\Pi_G^{gg}) \; + \; \frac{n_b}{n-1} u_m(\Pi_G^{gb}(n_b)) \right) \; \geq \; u_m(\Pi_G^{bb}) \tag{2}$$

where $p(n_G, t)$ is the probability that in period t there are n_G other type G players and $q_m(n_g, n_G)$ the probability that n_g of the type G players are less risk averse than player m.[14] Condition (2) compares for player m the expected utility of choosing *green* assuming all type G players with $x_i < x_m$ also choose *green* (left-hand side) with the utility when everyone chooses *blue* (right-hand side). Social change is initiated in the first period in which there exists a marginal player m for whom (2) holds.[15]

[13] For low disunity penalties, it can be efficient if type G players switch to *green* earlier than type B players: if $(n-1)p < v_B + v_G$, the efficient switching period is given by $\min\left(t : f_{G,t} \geq 1 - v_G/((n-1)p)\right)$ for type G and $\min\left(t : f_{G,t} \geq v_B/((n-1)p)\right)$ for type B.

[14] The probabilities are $p(n_G, t) = \binom{n-1}{n_G} f_{G,t}^{n_G} f_{B,t}^{n_B}$ and $q_m(n_g, n_G) = \binom{n_G}{n_g} \Phi\left(\frac{x_m - \mu}{\alpha}\right)^{n_g} \Phi\left(\frac{\mu - x_m}{\alpha}\right)^{n_G - n_g}$, Φ being the cumulative distribution function of the standard normal distribution.

[15] Because all players anticipate the deviation from the *blue* norm, in equilibrium the type B players as well as the remaining type G players switch to *green* in the same period as the players who are willing to initiate change (at least if p is sufficiently large, as is the case in our experiment).

Table 3: When Do Groups Switch to the *Green* Norm?

	Period of Change		
	Efficient	Risk Neutral	Elicited Risk Aversion
Baseline	8	12	No Change
High Return	4	6	19
Low Penalty	8	5	8
Endogenous Penalty[(a)]	8	9	16
Reward	8	4	19
Poll	8	12	14
Fast Information	8	12	No Change
Small Group	8	12	No Change

The efficient period of change and the predicted period of change under risk neutrality and for the elicited risk preferences (estimated via an interval regression based on the risk elicitation task) are based on the analytical framework, expressions (1) and (2).
(a) Penalty choices are assumed to be $p = 1$ for type G subjects and $p = 7$ for type B subjects.

5.4 Behavioral Hypotheses

Table 3 shows for the different treatments the socially efficient period of change, the predicted period of change assuming risk neutrality, and the predicted period of change when estimating the distribution of risk parameters using subjects' choices in the risk elicitation task.[16] The predictions are based on the analytical framework, specifically condition (2), which states that social change occurs as soon as there is a group of type G players who benefit from jointly deviating to *green*, even if the remaining players continued to choose *blue*.

Table 3 shows that when allowing for risk aversion (third column) in all treatments except Low Penalty social change occurs too late relative to the efficient outcome. In treatment Baseline, we predict a complete lock-in: change to *green* does not happen within the 31 periods.

Hypothesis 1: *In most treatments, change to green occurs later than in the socially efficient equilibrium.*

We next compare the different treatments (using the elicited risk preferences). Change is predicted to occur in treatments High Return, Low Penalty, and Endogenous Penalty. In Endogenous Penalty, we assume that type G players choose a low penalty ($p = 1$) to encourage change, while type B players choose large penalties ($p = 7$) to achieve the opposite. In period 16, when change is predicted to occur, the expected penalty against subjects choosing *green* equals $f_{B,t=16} * 7 + f_{G,t=16} * 1 = 2.24$. Note that in the presence of risk aversion, lowering potential punishments (cushioning the downside of change, i.e., the miscoordination costs) is more effective than increasing the returns to change.

[16] For the latter predictions, we assume constant absolute risk aversion (CARA) and estimate the distribution of risk parameters via an interval regression. For instance, a subject that chooses lottery (a) reveals that her CARA risk parameter is at least 0.79. A subject that chooses lottery (b) has a risk parameter in the interval $[0.33, 0.79]$, and so on. Assuming a normal distribution, interval regressions show that a mean of 0.15 and a standard error of 0.20 best describe the distribution of risk parameters.

In treatment Reward, the first four subjects initiating change receive the maximum earnings in their group. If subjects expect change to occur at some point in time there is an incentive to switch to *green* early to earn the reward. Once four subjects have switched to *green*, the remaining subjects face a different game with the incentives shifted toward choosing *green*. As a result, change is realized earlier than in the Baseline treatment.

In treatment Poll, before choosing colors in period 14, subjects have the possibility to anonymously (without incurring costs) reveal their preferences. This resolves uncertainty about the current number of type B and G players and, in addition, may provide a focal period in which change can occur. We expect that this suffices to induce change in period 14.[17]

Hypothesis 2: *Ranking treatments by the predicted period of change based on the elicited risk preferences, change occurs early in Low Penalty (period 8), followed by Poll (period 14), which in turn is followed by Endogenous Penalty (period 16), High Return, and Reward (both in period 19). Change fails to occur in Baseline, Fast Information, and Small Group.*

Finally, condition (2) implies that risk aversion delays social change, both a high mean μ and high standard deviation α. This is confirmed in table 3 when comparing the second and third column.

Hypothesis 3: *Groups are less likely to achieve social change if subjects exhibit a high degree of risk aversion. If change occurs, it is initiated by the least risk averse subjects.*

6 Experimental Results

We begin by discussing the main experimental results on the prevalence of the conformity trap in Section 6.1. We then discuss aspects related to efficiency in Section 6.2. Finally, in Section 6.3, we focus on the characteristics of initiators of change. Throughout the analysis the unit of observation for the non-parametric tests are session averages. If justified by the hypotheses derived in the previous section, we will report p-values from one-tailed tests.

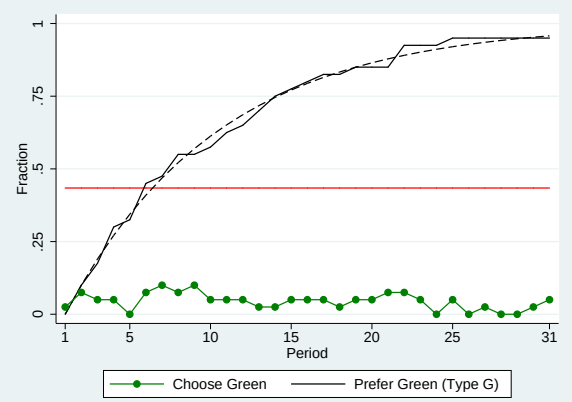
6.1 Social Change

Figure 2 depicts behavior in treatments Baseline and High Return. Figures (b) and (d) on the right-hand side show the behavior in each session. Figures (a) and (c) on the left-hand side summarize the six sessions in each treatment, each point corresponding to the median observation over the six sessions. The line with circled markers represents the fraction of subjects choosing *green* in each of the 31 periods. The solid line shows the fraction of type G subjects and the dashed line the corresponding theoretical expectation. The horizontal line depicts the fraction of subjects required to choose *green*,

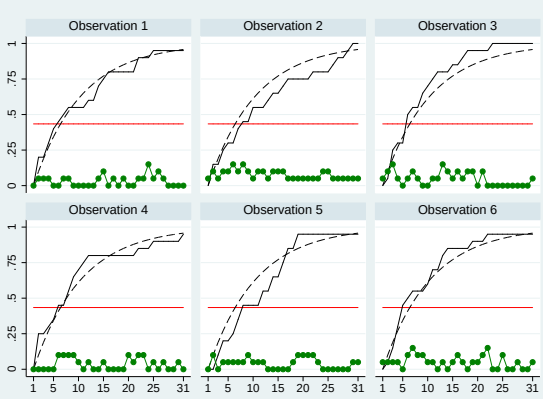
[17] The analytical framework does not allow us to differentiate predictions for Fast Information, Small Group, and the Baseline treatment. Empirically, however, our expectation is that accelerating feedback in Fast Information and a smaller group size facilitate coordination and change.

Figure 2: Color Choices in Baseline and High Return

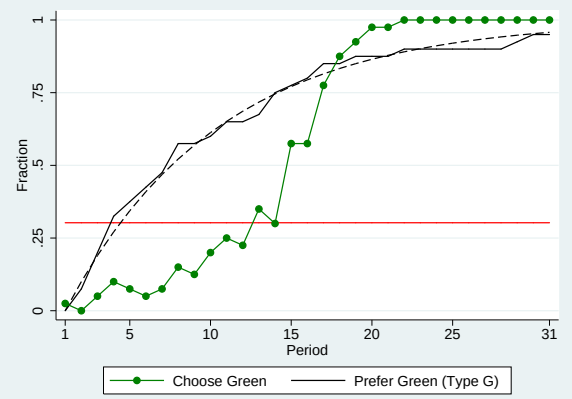
(a) Baseline



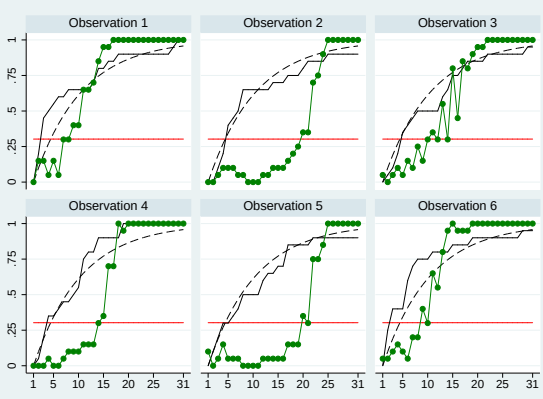
(b) Baseline: Sessions



(c) High Return



(d) High Return: Sessions



The line with circled markers represents the fraction of subjects choosing *green* in each period. The solid increasing line shows the fraction of type G subjects; the dashed line the corresponding theoretical expectation. The horizontal line is the tipping point (the fraction of subjects choosing *green* such that a risk neutral type G subject prefers to choose *green* as well). Figures (a) and (c) depict the median observation over the six sessions shown in Figures (b) and (d).

such that a risk neutral type G subject prefers to choose *green* as well. It can be interpreted as a tipping point. If groups reach it, they are unlikely to fall back to the *blue* norm.

Result 1 (Baseline)**:** *All groups are caught in the conformity trap, even though it is commonly known that virtually all subjects prefer green.*

Support: The results in Baseline provide clear evidence of the conformity trap. As can be seen in Figure 2b, groups in all six sessions play *blue* until and including the last period, even though on average the majority of subjects was type G after period 8 and almost every subject was type G by the end of the experiment. In each session there were periods in which several participants chose *green* (up to three at the same time), but these attempts at triggering change were never successful. □

Result 2 (High Return)**:** *Higher returns for type G subjects when choosing green facilitate change,*

although change occurs much later than in the efficient outcome.

Support: Figures 2c and 2d depict the results for High Return. In all six sessions groups broke out of the conformity trap, significantly different from the Baseline (one-sided Fisher's exact test, p=0.001). Notice that change tends to be slow until the tipping point is reached and speeds up thereafter. The periods in which the tipping point was reached are 7, 20, 10, 14, 20 and 9, substantially later than in the socially efficient equilibrium where the switch to the *green* norm occurs already in period 4. Indeed, groups would have been better off choosing *green* already in the first period, avoiding being trapped at the *blue* norm. □

The results in Baseline establish the conformity trap as a real phenomenon. Treatment High Return shows that groups can escape the conformity trap, but with a substantial delay relative to the first-best outcome. We next turn to some interventions designed to encourage individuals to lead change. In Low Penalty, the cost to non-conformity are reduced. In Endogenous Penalty, subjects choose the disunity penalties themselves. In Reward, disunity penalties are as in the Baseline treatment, but leaders (the first ones to initiate change) are rewarded.

Result 3 (Low Penalty)**:** *Lower disunity penalties facilitate change.*

Support: The results for Low Penalty in Figure 3a and 3b show that lowering the disunity penalty parameter to $p = 1$ mitigates the conformity trap. In five out of six sessions, groups managed to adapt to the *green* norm. The difference to Baseline is significant (one-sided Fisher's exact test, p=0.008). Despite the low penalties, however, the effects of conformity are still visible. Efficiency would require subjects to switch to *green* by period 8, but change did not occur in session 1 and occurred with delay in all other sessions. □

Result 7.1 (Color Choices in Endogenous Penalty)**:** *The likelihood of groups getting caught in the conformity trap in Endogenous Penalty is not lower than in Baseline.*

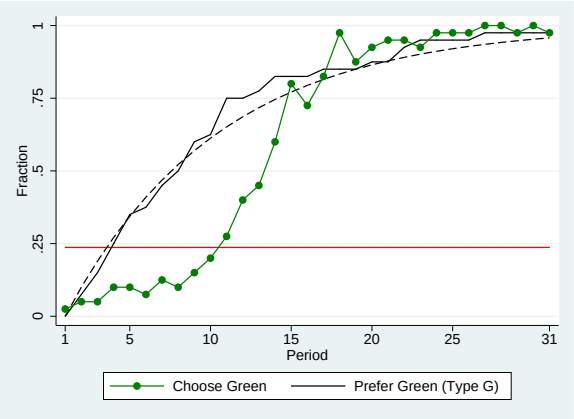
Support: Five out of six groups in Endogenous Penalty were unable to escape the conformity trap, see Figure 3d. Social change is thus not significantly more likely than in Baseline (one-sided Fisher's exact test, p=0.500), and this despite the option to choose $p = 1$, which corresponds to the value in Low Penalty in which change was frequently observed. Letting subjects choose penalties does not facilitate change in the social change game. The regressions in Table 4 presented in Section 6.3 will show that subjects were more likely to deviate from the *blue* norm than in Baseline, but this was not sufficient to trigger change (except in one session). □

Result 7.2 (Penalty Choices in Endogenous Penalty)**:** *Penalty choices of subjects choosing blue (i.e., penalties against green) increase over time and do not differ by type B and G.*

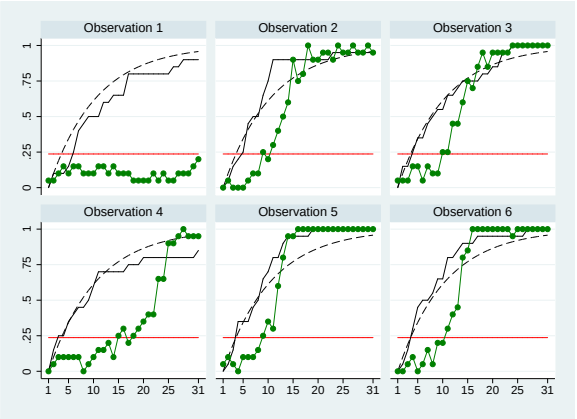
Support: The penalty choices in Endogenous Penalty are shown in Figure 4. The figure depicts the fraction of low ($p = 1$), medium ($p = 4$) and large ($p = 7$) penalty choices of subjects choosing *blue*. The total height of each bar thus corresponds to the fraction of subjects choosing *blue*. The solid line depicts their average penalty choice, which is below $p = 4$ in most periods. Interestingly, there is a significant upward trend in average penalties (Spearman's rho above 0.7). This is true irrespective of

Figure 3: Color Choices in Low Penalty, Endogenous Penalty, and Reward

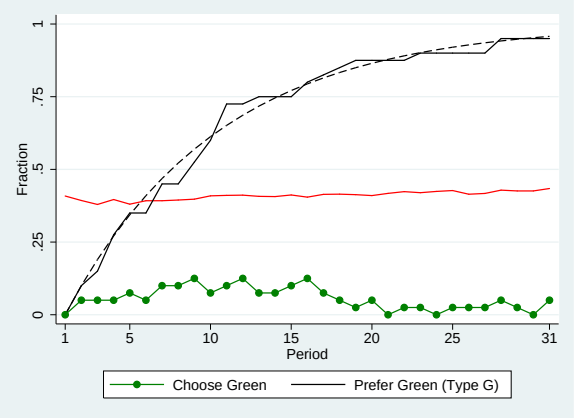
(a) Low Penalty



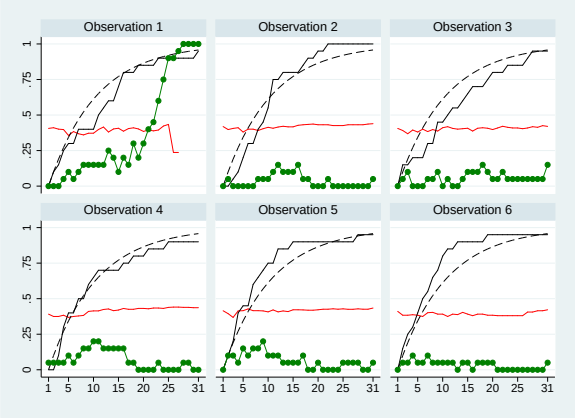
(b) Low Penalty: Sessions



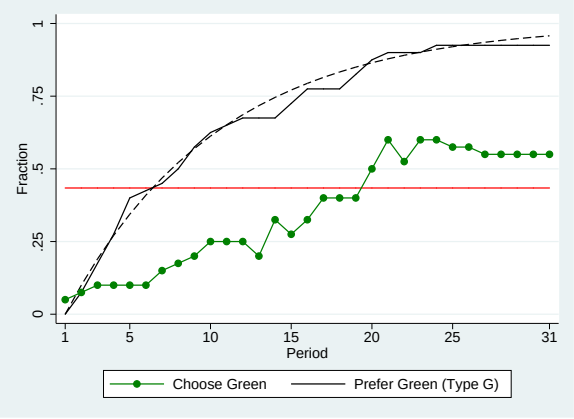
(c) Endogenous Penalty



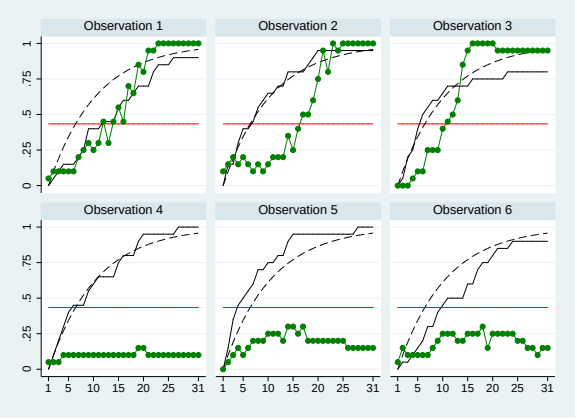
(d) Endogenous Penalty: Sessions



(e) Reward



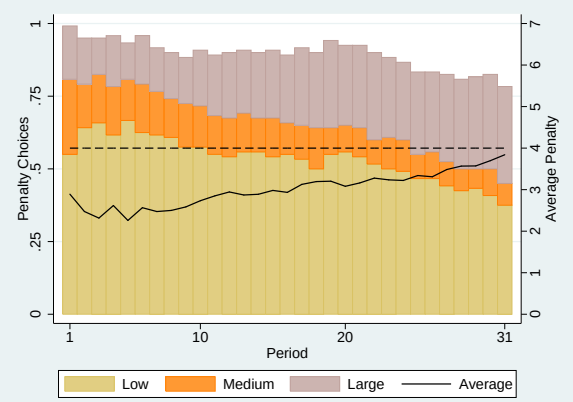
(f) Reward: Sessions



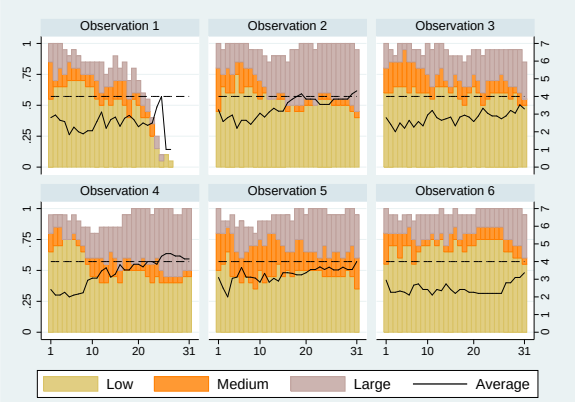
The line with circled markers represents the fraction of subjects choosing *green* in each period. The solid increasing line shows the fraction of type G subjects; the dashed line the corresponding theoretical expectation. The horizontal line is the tipping point (the fraction of subjects choosing *green* such that a risk neutral type G subject prefers to choose *green* as well). Figures (a), (c) and (e) depict the median observation over the six sessions shown in Figures (b), (d) and (f).

Figure 4: Penalty Choices of Subjects Choosing *blue*

(a) Penalty Choices



(b) Penalty Choices: Sessions



Fraction of low ($p = 1$), medium ($p = 4$) and large ($p = 7$) penalty choices and corresponding average penalty (solid line) of subjects choosing *blue* (i.e., penalties against *green*). The total height of each bar corresponds to the fraction of subjects choosing *blue*.

subjects' types, i.e., it is true even for subjects who prefer *green*. In fact, the average penalty selected by type G subjects choosing *blue* was 3.12, slightly *higher* than the average penalty of 2.74 chosen by type B subjects (Wilcoxon matched-pairs signed-ranks test, $p = 0.66$). This highlights an interesting trade-off: if disunity penalties are low, change is more likely to occur, but due to the uncertainty about whether change will in fact be realized, subjects also have an incentive to choose high penalties to avoid costly miscoordination. □

Result 5 (Reward)**:** *Rewards for initiators of change help groups adapt to the green norm but the conformity trap is still common.*

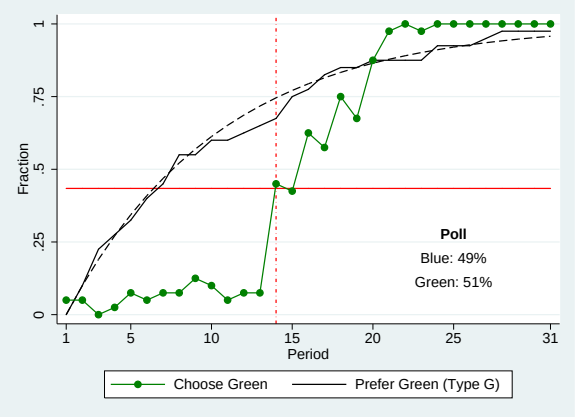
Support: Figures 3e and 3f summarize behavior in treatment Reward. Three groups were caught in the conformity trap and three groups escaped it. The difference to Baseline is (weakly) significant (one-sided Fisher's exact test, p=0.090). The regressions presented in Section 6.3 (Table 4) will confirm that subjects in Reward were more likely to try instigating change than in Baseline. Notice that in sessions 4, 5 and 6 several subjects chose *green* early on but eventually failed to induce change. Trying to instigate change thus involves the risk of incurring large losses. This is illustrated by the low average payoffs of 15.68 in sessions where change failed to occur. These payoffs are substantially lower than the average payoff in Baseline of 19.93 (Mann-Whitney U test, p=0.020). Interestingly, providing rewards does promote the emergence of leaders as predicted, but it sometimes does only that, failing to incentivise the "marginal" subject required to reach the tipping point. □

Finally, we look at treatments Poll, Fast Information and Small Group. These treatments are expected to promote social change by helping group members coordinate and synchronize their efforts to break out of the conformity trap.

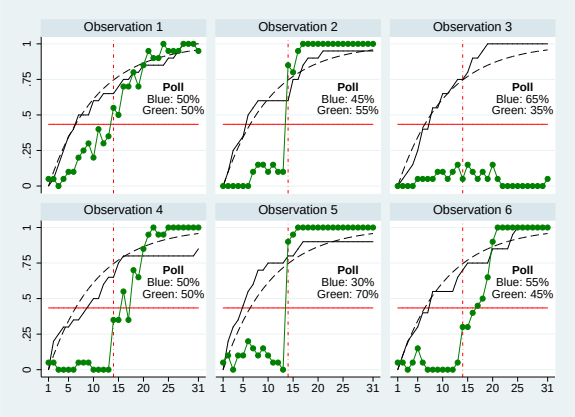
Result 6 (Poll)**:** *Most groups use the poll to break out of the conformity trap.*

Figure 5: Color Choices in Poll, Fast Information and Small Group

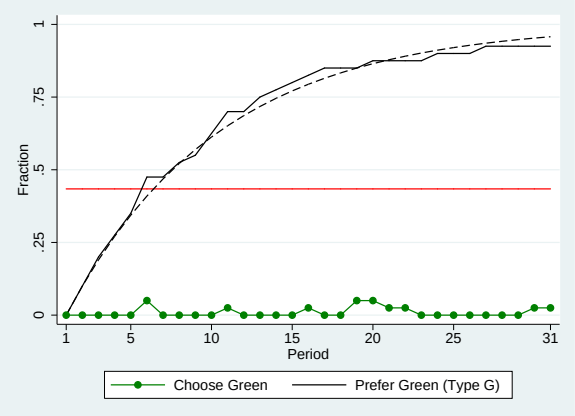
(a) Poll



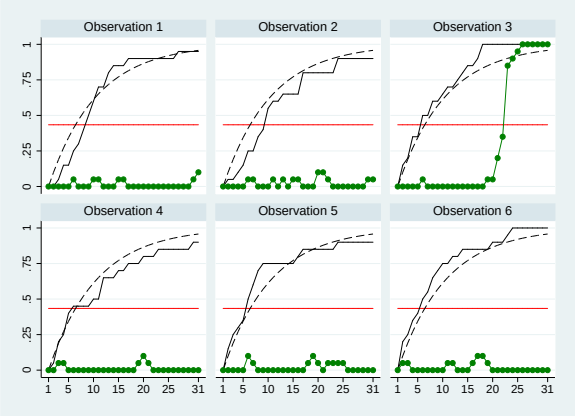
(b) Poll: Sessions



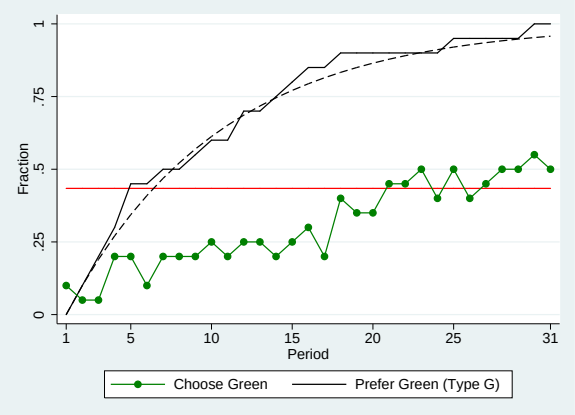
(c) Fast Information



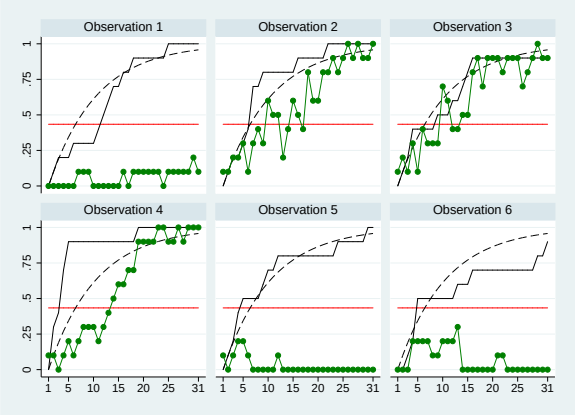
(d) Fast Information: Sessions



(e) Small Group



(f) Small Group: Sessions



The line with circled markers represents the fraction of subjects choosing *green* in each period. The solid increasing line shows the fraction of type G subjects; the dashed line the corresponding theoretical expectation. The horizontal line is the tipping point (the fraction of subjects choosing *green* such that a risk neutral type G subject prefers to choose *green* as well). Figures (a), (c) and (e) depict the median observation over the six sessions shown in Figures (b), (d) and (f). The vertical line in Poll indicates period 14.

Support: The results for Poll are presented in Figure 5a and 5b. The figure includes as additional information the percentage of subjects who stated *blue* or *green* as their preferred color in the poll in period 14. Polls are effective coordination devices in our experiment, leading to social change in five out of six sessions, a significant difference to Baseline (one-sided Fisher's exact test, p=0.008). Change usually occurred in the period in which the poll was conducted, or shortly thereafter. □

The group in session 3 of the Poll treatment did not manage to escape the conformity trap. The poll result shows that even though more than 75% of the participants were type G in period 14, only 35% stated that they prefer others to choose *green* in the next rounds. The observation that not all type G participants state *green* as their preferred color is a robust finding: while type B subjects voted *blue* in most cases (86%), type G subjects voted *green* only 67% of the time (Wilcoxon matched-pairs signed-ranks test, p=0.074), explaining why the poll was split 50-50 on average. Why are type G subjects reluctant to vote for *green*? Looking at observations 4 and 6, we see that when the poll was split, it was followed by a phase of disagreement. In other words, the poll leads to a quick and efficient change only if a substantial majority votes for *green*. If subjects believe that the majority won't be sufficiently large to avoid miscoordination, they may vote for *blue* despite the fact that the poll is anonymous and preferences can be revealed without incurring disunity penalties.

Result 7 (Fast Information)**:** *Expediting feedback does not reduce the likelihood of groups getting caught in the conformity trap compared to Baseline.*

Support: Figures 5c and 5d summarize behavior in Fast Information. Social change occurred in only one out of six sessions, not significantly different from Baseline (one-sided Fisher's exact test, p=0.500). However, providing immediate feedback about the color choices of everyone in the group did affect behavior. Subjects in Fast Information were on average *less* likely to choose *green* than participants in Baseline, as is apparent from the fact that the line depicting the fraction of subjects choosing *green* is flatter than in Baseline. The regressions presented in the Section 6.3 (Table 4) show that the reduction in the willingness to choose *green* (when most others still choose *blue*) is statistically significant. At the same time, behavior in session 3 suggests that immediate feedback can also expedite change once it has been initiated. □

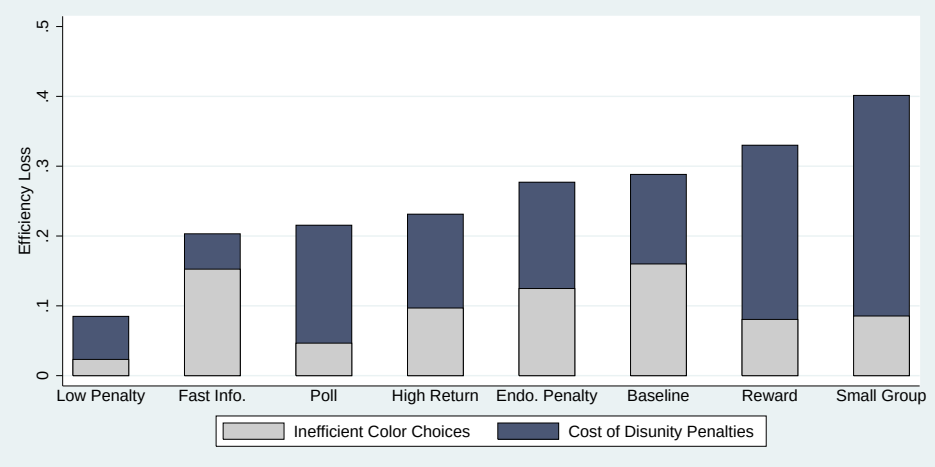
Result 8 (Small Group)**:** *Smaller groups are more likely to escape the conformity trap but tend to incur large disunity costs due to miscoordination.*

Support: Figures 5e and 5f depict the results for Small Group: three groups were caught in the conformity trap and three groups escaped it, a (weakly) significant difference to Baseline (one-sided Fisher's exact test, p=0.090).[18] In the sessions where change was achieved, the transitions from *blue* to *green* involved prolonged phases during which subjects failed to coordinate on one color. As a consequence, the average per period payoffs realized in Small Group in the sessions where change was achieved (sessions 2, 3 and 4) was only 13.74, substantially lower than the average payoff in the sessions without change (19.79). □

In sum, we find clear support for Hypothesis 1 (delayed social change relative to the efficient behavior).

[18]Figure 5e should be interpreted with caution. Since behavior in Small Group was split, the median is computed as the average of the behavior in two sessions.

Figure 6: Efficiency Loss Relative to Social Optimum



Efficiency loss relative to the socially efficient outcome. The gray (lower) part of each bar shows the inefficiency due to color choices; the darker (upper) part the inefficiency due to disunity penalties.

In Hypothesis 2, we have ranked the different treatments based on the predicted period of change. The qualitative predictions were largely correct. In particular, reducing disunity penalties is by far the most effective intervention to promote change. Strikingly, however, we also saw that allowing subjects to choose the penalties themselves does not suffice to promote change: in treatment Endogenous Penalty subjects chose large penalties to avoid miscoordination even when they preferred *green*. Another intervention that was less effective in expediting change than hypothesized is the promise of rewards to leaders of social change, interestingly not because of a lack of leaders but their failure to attract enough followers.

6.2 Efficiency

Results 1 to 8 document the prevalence of the conformity trap. Groups almost always fail to switch to *green* before or at the efficient period, and often change is not achieved even by the final period. What are some implications for social welfare? In Figure 6, we depict the efficiency loss in each treatment relative to the socially efficient outcome. To that end, we divided the total realized earnings by the earnings subjects would have made if they had played according to the socially efficient outcome. The light gray bars show the efficiency losses due to inefficient color choices; the darker part of each bar the efficiency loss due to disunity penalties.

Result 9 (Efficiency)**:** *Efficiency losses relative to the social optimum are substantial and are caused by both disunity penalties and delayed (or absence of) change.*

Support: Figure 6 shows that efficiency losses range from 8% in Low Penalty to 29% in Baseline and 40% in Small Group. Averaged over all treatments and periods, 39% of the efficiency loss is due to delayed or no social change (i.e., not choosing the preferred color), while 61% is due to miscoordination (i.e., disunity penalties). Thus, both sources of inefficiency are important. Low Penalty is

significantly more efficient than all other treatments (Mann-Whitney U test, p=0.004 when compared to Baseline). Fast Information and High Return are more efficient than Baseline (p=0.004). The differences in efficiency between Baseline and the three conditions Poll / Endogenous Penalty / Reward are insignificant (p=0.150 / p=0.423 / p=0.873).[19] The least efficient treatment is Small Group (p=0.078 if compared to Baseline). □

6.3 Initiators of Change

Table 4 provides insights about the attributes of initiators of change, the first individuals to defy the *blue* norm. The dependent variable of the regressions is the probability of choosing *green* in periods in which the percentage of subjects choosing *green* was at most 20%. The dummy *risk-accepting* is generated using the risk elicitation task (see Section 5.2), equalling 0 for subjects who chose lotteries that imply risk-aversion and 1 otherwise. We also elicited subjects' tendency for non-conformity. If the score of a subject in the survey exceeded the median score among the 900 participants, we classify her as a *non-conformist*.[20] We find significant correlations between the probability of initiating change with both the risk and the non-conformity measure.[21]

Result 10 (Risk and Non-Conformity)**:** *Initiators of change tend to have more tolerance for risk and a greater dislike for conformity.*

Support: Table 4 shows that risk-accepting subjects were more likely to be initiators of change. Their probability of choosing *green* is 1.6 percentage points higher. Notice that risk preferences should mainly affect individuals' decision if they take into account how it might affect future outcomes, indicating the presence of forward-looking subjects. The coefficient of the non-conformity measure is also significant, implying that some subjects derive utility from non-conformity and from leading change. □

The results in Table 4 also confirm that most treatments lead to more attempts at initiating change than witnessed in the Baseline treatment. The coefficient for treatment Poll is insignificant, but recall that there the poll functions as a coordination device and groups depend to a lesser degree on individual deviations to achieve change. Interestingly, in Fast Information attempts at choosing *green* are significantly *less* common than in Baseline, suggesting that expediting social feedback may promote conformity rather than helping groups coordinate on change.

We also ran a similar regression only for treatment Reward and found the effect of the risk dummy to be accentuated. Not surprisingly, the subjects who deviated first to receive the reward were the least

[19] Notice that most observations in Poll are substantially more efficient than each of the six sessions in Baseline (the p-value of 0.150 is due to Session 1 and 3, which are slightly less efficient than the average Baseline session). Poll is not significantly less efficient than Fast Information (Mann-Whitney U test, p=1.000).

[20] A subject's score for each statement in the conformity task was determined on a five-point rating scale from 1 ("strongly disagree") to 5 ("strongly agree"). The total score corresponds to the sum of the scores for the ten statements subjects had to consider.

[21] The regression results are robust to using as independent variables the six lottery choices and the score of the conformity measure rather than the dummy variables. We also included political ideology (Democrat versus Republican) and ethnicity but these variables were insignificant and did not affect the other coefficients. The results are also robust to using the tipping point or 30% choosing *green* as the cut-off fraction of *green* choices.

Table 4: Determinants of the Probability of Choosing *green*

Prob(Green|θ = G)	(1)	(2)	(3)	(4)
High Return	0.025***	0.024***	0.024***	0.024***
	(0.006)	(0.006)	(0.006)	(0.006)
Low Penalty	0.041***	0.041***	0.042***	0.042***
	(0.006)	(0.007)	(0.006)	(0.006)
Endogenous Penalty	0.015**	0.015**	0.014**	0.014**
	(0.007)	(0.007)	(0.007)	(0.007)
Reward	0.023***	0.024***	0.023***	0.023***
	(0.009)	(0.009)	(0.009)	(0.009)
Poll	0.010	0.010	0.010	0.010
	(0.008)	(0.008)	(0.009)	(0.009)
Fast Information	-0.019***	-0.020***	-0.021***	-0.021***
	(0.005)	(0.005)	(0.005)	(0.005)
Small Group	0.024**	0.023**	0.023**	0.023**
	(0.010)	(0.009)	(0.010)	(0.010)
Risk-Accepting	0.016***		0.016***	0.016***
	(0.006)		(0.006)	(0.006)
Non-Conformist		0.015***	0.015***	0.014***
		(0.005)	(0.005)	(0.005)
Female				-0.005
				(0.005)
Period Dummies	Yes	Yes	Yes	Yes
Fraction of *green*	≤ 0.20	≤ 0.20	≤ 0.20	≤ 0.20
Observations (Individuals)	11,363 (643)	11,363 (643)	11,363 (643)	11,363 (643)
Clusters (Sessions)	48	48	48	48

* p<0.1; ** p<0.05; *** p<0.01. Average marginal effects of mixed effects probit regressions with session and individual random intercepts. Standard errors clustered on sessions. Observations include type G subjects in periods in which less than 20% chose *green*. The reference treatment is Baseline.

risk averse ones. Because these are also the individuals most likely to deviate first in the absence of a reward, the rewards were less effective in bringing change than hypothesized. This poses the question whether in daily life societies tend to undervalue the contributions of key followers compared to those of leaders.

7 Concluding Remarks

Our study provides evidence that social norms and institutions can fail to adapt in a changing world when there is a strong pressure to conform. Remarkably, this can occur even when it is common knowledge among participants that social change would be widely beneficial for them. These findings suggest that rules adopted by societies to promote welfare at one point, may end up having the opposite effect in the long-run by hindering change. Coordination problems have been previously identified as one possible reason why socially beneficial change may fail, but usually have been discussed in combination with other issues hindering change such as conflicting interests or the presence of powerful elites (e.g., Arthur, 1989; North, 1990; Acemoglu et al., 2005). Our study presents evidence that coordination problems alone can lead to social or institutional stagnation.

Facilitating change, even when it is widely beneficial for society, is a non-trivial task. Several of our

interventions had only a limited impact, and others had none. What seems clear from our research is the role of leadership. One person or coordinated group of first-movers must be willing to suffer extreme losses in the short run in order to lower the cost for the next group, and so on until a cascade of change comes. A second and related lesson is the role played by luck in getting change. Many people may stick their necks out when it is privately recognized that change should, by efficiency standards, be coming. But only if by chance enough others also move toward change at the same time can the cascading described above be witnessed. A third lesson is the important role of hope. In our conditions with "public opinion polls," the good news is that rapid change will come after a majority reveals to the pollster that they prefer change. Surprisingly, however, many times subjects did not actually register this view with the pollster. This conservatism can best be understood as pessimism about the process of change and the miscoordination cost associated with it. Similarly, in a different condition where we provide immediate and complete information about group behavior at the end of each round, subjects are less likely to try instigate change. The constant and immediate reminder that everyone is following the norm seems to spread pessimism.

We view our study as a starting point to a research program investigating empirically the impact different forces have on social change. Our experimental paradigm can be easily extended to explore richer environments, for instance when individuals interact in social networks. Empirical investigation is essential given the multiple equilibria arising whenever there is a need for coordination. On the theoretical side, the literature on global games and dynamic coordination games more generally (e.g., Angeletos et al., 2007; De Mesquita, 2010) delivers important insights about the role of information in promoting or preventing change. Our findings complement these studies, emphasizing the importance of alleviating the dilemma faced by pioneers of change. In a rapidly changing world, societies need to ensure that progress does not depend only on the "unreasonable man" who is willing to suffer large costs to promote social change.

References

Acemoglu, Daron and James A Robinson, "Persistence of power, elites, and institutions," *American Economic Review*, 2008, *98* (1), 267–93.

_ **and Matthew O. Jackson**, "History, expectations, and leadership in the evolution of social norms," *The Review of Economic Studies*, 2015, *82* (2), 423–456.

_ **, Simon Johnson, and James A Robinson**, "Institutions as a fundamental cause of long-run growth," *Handbook of economic growth*, 2005, *1*, 385–472.

Angeletos, George-Marios, Christian Hellwig, and Alessandro Pavan, "Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks," *Econometrica*, 2007, *75* (3), 711–756.

Arthur, W. Brian, "Competing technologies, increasing returns, and lock-in by historical events," *The Economic Journal*, 1989, *99* (394), 116–131.

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

_ **and Ryan Muldoon**, "Social Norms," *Stanford Encyclopedia of Philosophy*, 2011.

Brandts, Jordi and David J. Cooper, "A change would do you good.... An experimental study on how to overcome coordination failure in organizations," *The American Economic Review*, 2006, *96* (3), 669–693.

Brock, William A. and Steven N. Durlauf, "Discrete choice with social interactions," *The Review of Economic Studies*, 2001, *68* (2), 235–260.

Coleman, James S., *Foundations of social theory*, Harvard University Press, 1994.

David, Paul A., "Clio and the Economics of QWERTY," *The American Economic Review*, 1985, *75* (2), 332–337.

De Mesquita, Ethan Bueno, "Regime Change and Revolutionary Entrepreneurs," *American Political Science Review*, 2010, *104* (3).

Elster, Jon, "Social norms and economic theory," *Journal of Economic Perspectives*, 1989, *3* (4), 99–117.

Farrell, Joseph and Garth Saloner, "Standardization, compatibility, and innovation," *The RAND Journal of Economics*, 1985, pp. 70–83.

Fehr, Ernst and Simon Gächter, "Cooperation and punishment in public goods experiments," *American Economic Review*, 2000, *90* (4), 980–994.

Fischbacher, Urs, "z-Tree: Zurich toolbox for ready-made economic experiments," *Experimental Economics*, 2007, *10* (2), 171–178.

Friedman, Daniel, "Equilibrium in evolutionary games: Some experimental results," *The Economic Journal*, 1996, pp. 1–25.

Gërxhani, Klarita and Jeroen Bruggeman, "Time lag and communication in changing unpopular norms," *PloS one*, 2015, *10* (4), e0124715.

Goldsmith, Ronald E., Ronald A. Clark, and Barbara A. Lafferty, "Tendency to conform: a new measure and its relationship to psychological reactance," *Psychological reports*, 2005, *96* (3), 591–594.

Greif, Avner and David D. Laitin, "A theory of endogenous institutional change," *American Political Science Review*, 2004, *98* (04), 633–652.

Harsanyi, John C. and Reinhard Selten, *A general theory of equilibrium selection in games*, Vol. 1, The MIT Press, 1988.

Heggedal, Tom-Reiel and Leif Helland, "Platform selection in the lab," *Journal of Economic Behavior & Organization*, 2014, *99*, 168–177.

Hong, Sung-Mook and Salvatora Faedda, "Refinement of the Hong psychological reactance scale," *Educational and Psychological Measurement*, 1996, *56* (1), 173–182.

Hossain, Tanjim and John Morgan, "The quest for QWERTY," *The American Economic Review Papers and Proceedings*, 2009, *99* (2), 435–440.

_ , Dylan Minor, and John Morgan, "Competing matchmakers: An experimental analysis," *Management Science*, 2011, *57* (11), 1913–1925.

Kandori, Michihiro, George J. Mailath, and Rafael Rob, "Learning, mutation, and long run equilibria in games," *Econometrica*, 1993, pp. 29–56.

Katz, Michael L. and Carl Shapiro, "Network externalities, competition, and compatibility," *The American Economic Review*, 1985, *75* (3), 424–440.

Liebowitz, Stan J. and Stephen E. Margolis, "Network externality: An uncommon tragedy," *The Journal of Economic Perspectives*, 1994, *8* (2), 133–150.

_ and _ , "Path dependence, lock-in, and history," *Journal of Law, Economics, & Organization*, 1995, *11*, 205.

Mackie, Gerry, "Ending footbinding and infibulation: A convention account," *American Sociological Review*, 1996, *61* (6), 999–1017.

Masiliūnas, Aidas, "Overcoming coordination failure in a critical mass game: Strategic motives and action disclosure," *Journal of Economic Behavior & Organization*, 2017, *139*, 214–251.

Michaeli, Moti and Daniel Spiro, "From Peer Pressure to Biased Norms," *American Economic Journal: Microeconomics*, 2017, *9* (1), 152–216.

North, Douglass C., *Institutions, institutional change and economic performance*, Cambridge university press, 1990.

Ostrom, Elinor, "Collective action and the evolution of social norms," *Journal of economic perspectives*, 2000, *14* (3), 137–158.

Peski, Marcin, "Generalized risk-dominance and asymmetric dynamics," *Journal of Economic Theory*, 2010, *145* (1), 216–248.

Shaw, George Bernard, *Man and superman: a comedy and a philosophy*, Brentano's, 1903.

Smerdon, David, Theo Offerman, and Uri Gneezy, "Everybody's doing it: On the Emergence and Persistence of Bad Social Norms," *Tinbergen Institute Discussion Paper No. 16-023/I*, 2016.

Van Huyck, John B., Raymond C. Battalio, and Richard O. Beil, "Tacit coordination games, strategic uncertainty, and coordination failure," *The American Economic Review*, 1990, *80* (1), 234–248.

Weinstein, Jay, *Social change*, Rowman & Littlefield Publishers, 2010.

Wilkening, Tom, "Information and the persistence of private-order contract enforcement institutions: An experimental analysis," *European Economic Review*, 2016, *89*, 193–215.

Williamson, Oliver E., "The new institutional economics: taking stock, looking ahead," *Journal of Economic Literature*, 2000, *38* (3), 595–613.

Young, Peyton, "The evolution of conventions," *Econometrica*, 1993, *61* (1), 57–84.