Social change and the conformity trap[*]

James Andreoni[‡] Nikos Nikiforakis[§] Simon Siegenthaler[¶]

July 1, 2019

First draft: 20. November 2016

Abstract

The ability of societies to adapt their norms to changing circumstances is critical for welfare. We study a game in which individuals initially coordinate on a mutually preferred norm but, as preferences gradually change, switching to a different equilibrium becomes desirable for everyone in the group. However, transitioning to the newly preferred equilibrium is challenging due to a pressure to conform to the action chosen by the majority. We present evidence from a laboratory experiment showing that groups often fail to adapt their behavior and explore factors that may help avoid such conformity traps. While several interventions facilitate change—e.g. reducing the pressure to conform, providing more precise information about the speed at which preferences change, or aggregating beliefs via an opinion poll—the key to beneficial adaptation are individuals who are committed to change despite incurring large losses in payoff.

JEL Classification: C92, D60, D70

Keywords: Conformity, Coordination Failure, History Dependence, Network Effects, Social Norms

[*] We would like to thank Vincent Leah-Martin, Seung-Keun Martinez and Alex Kellogg for excellent research assistance. We thank Olivier Bochet, Aurélie Dariel, Nisvan Erkal, Guillaume Fréchette, Marie Claire Villeval, Thomas Palfrey, Ariel Rubinstein, 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, the 2018 ASSA meetings in Philadelphia and seminar participants at the Wissenschaftszentrum Berlin für Sozialforschung as well as the universities of Texas at Dallas, 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: +971 262-85436.

[¶] Naveen Jindal School of Management, University of Texas at Dallas, Richardson, TX 75080, USA. E-mail: simon.siegenthaler@utdallas.edu. Phone: +1 972-883-5871.

> *"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

Social norms can be defined as widely-accepted rules of behavior governing interactions among people, prescribing how one ought to behave and indicating what actions ought to be sanctioned (e.g., Bicchieri, 2006; Young, 2008). These rules influence our decisions in a broad variety of circumstances ranging from the determination of property rights and workplace etiquette, to behavior in public spaces and obligations towards strangers, colleagues and family alike (e.g., Elster, 1989; Posner, 2000; Williamson, 2000). Economists' interest in social norms is explained by the observation that they can enhance efficiency by reducing externalities or transaction costs (e.g., Arrow, 1971; Akerlof, 1976; Young, 2015). A fundamental question regarding social norms is how they change over time and whether change occurs when it is socially beneficial (e.g., Weinstein, 2010; Bicchieri, 2017).

Research across disciplines has emphasized that norms and standards, once established, can be slow to change and adapt to new circumstances—see e.g., Schelling (1978) and North (1990) in economics, Ostrom (2000) and Greif and Laitin (2004) in political economy, and Coleman (1990) and Centola et al. (2018) in sociology. Examples of norms discussed in this context include bans on interracial and same-sex marriage, popular acceptance of tobacco use, gender roles in the workplace, female labor participation, child marriage, footbinding, female genital cutting, and norms of personal retribution (Boehm, 1984; Elster, 1989, 1990; Bikhchandani et al., 1992; Coleman, 1994; Mackie, 1996; Dahlerup and Freidenvall, 2005; Bicchieri and Muldoon, 2011; Fernández, 2013; Bicchieri, 2017; Bursztyn et al., 2018). Another example is the norm to remain quiet about sexually abusive behavior, existent within specific groups as recently exposed in the film industry by what has become known as the #MeToo movement.[1]

In this paper, we contribute to the question of what can cause entrenchment of unproductive or inefficient paradigms, like in the examples above, and what forces can spark and accelerate widely beneficial change. The reasons that can potentially explain slow social change are diverse. They include incomplete or imperfect information (e.g., Bikhchandani et al., 1992; Kuran, 1995), expectations where increased prevalence of an institution enhances beliefs of future prevalence (e.g., Arthur, 1989; Greif and Laitin, 2004), the role of powerful elites in preventing change (e.g., North, 1990; Acemoglu and Robinson, 2008), and coordination effects where the value of a social institution or norm depends on the number

[1] We are interested in these examples, because they involve situations in which a majority would like to see change, and would have the power to achieve it, but change is nonetheless slow to materialize and depends on individuals who are willing to incur the costs of leading change. In the case of the #MeToo movement, 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. Yet, for a long time no one was willing to come forward out of fear of professional retribution, also not those who were not directly affected but knew about Weinstein's crimes. Then, on October 5, the *New York Times* used both legal documents and personal interviews with now famous actresses 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. 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.

of other people participating in it (e.g., Oliver et al., 1985; Kandori et al., 1993; Young, 1993; Acemoglu and Jackson, 2015).[2] In this article, we highlight the coordination aspect of social change, and we also address issues related to information and expectations. On the other hand, we abstract from the role of powerful elites. Indeed, we are interested in situations where agents are equally powerful and where, at least after some point in time, everyone in a group has the same preference over outcomes. Nevertheless, there is no guarantee that actual behavior will be in line with the group's preference.

We propose a new experimental paradigm, henceforth referred to as the social change game, capturing the tension between the need to coordinate and the need for change, and how this tension is mediated by beliefs, information and a range of institutional parameters. The need to coordinate is a central feature in many social interactions whether they involve social norms, customs or conventions and miscoordination entails individual costs, either due to social sanctions against individuals who fail to conform with the norm (e.g., Ostrom et al., 1992; Fehr and Gächter, 2000; Andreoni et al., 2003; Cinyabuguma et al., 2005; Balafoutas et al., 2014) or social image, feelings of guilt and not belonging (e.g., Asch, 1956; Jones, 1984; Bernheim, 1994; Akerlof and Kranton, 2000; Andreoni and Bernheim, 2009; Hopfensitz and Reuben, 2009; Goeree and Yariv, 2015; Bursztyn et al., 2017). The need for change arises as preferences evolve over time reflecting the arrival of new information. In such an environment, "pioneers of 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.

A general description of the social change game is as follows. There is a set of players which belongs to a large group. Players interact over multiple rounds. 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," which is increasing in the number of people in the group choosing the opposite color (capturing the need to coordinate). Thus, everyone choosing *blue* and everyone choosing *green* are both equilibria of the one-period game. Preferences exogenously change over time at a known rate, reflecting the arrival of new information. Specifically, 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* (capturing the need for a change in behavior). All aspects of the game are common knowledge. This game has a large number of Perfect Bayesian equilibria. The most natural ones involve all players choosing *blue* in the first periods followed by a switch to *green* once the number of players preferring *green* is sufficiently large.

The assumption that players are matched in pairs and incur costs only when failing to coordinate introduces an important non-linearity. If a single players deviates from an established norm, she will be matched with someone who is still adhering to the established norm. As a result, she will miscoordinate and incur a large cost since everyone else acts differently than her. However, as players gradually switch behavior to the alternative norm, the next players to deviate will not only receive a lower cost in case of miscoordination, but they will also be increasingly more likely to be matched with someone who already supports the new norm, in which case this pair of players incurs no costs. Using terminology

[2] Coordination and network effects 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).

borrowed from Oliver et al. (1985), this is situation of collective action with an "accelerating production function," characterized by daunting start-up costs for the initiators of social change but once a critical mass of initial deviations has materialized, rapid and complete change follows.[3]

To formulate hypotheses about likely outcomes in the social change game, we conceptualize the tension between the old norm (*blue*) and the new norm (*green*) as a war of attrition, adapting a model by Egorov and Harstad (2017).[4] Players who prefer *blue* want to delay change, while players whose preferences have already switched to *green* are in favor of immediate change. The war of attrition arises because each side privately knows how well it can deal with miscoordination and thus how long it will be able to persist in choosing the preferred color before conceding to the other side. A new feature in our model is that, because players' preferences gradually change to *green*, the likelihood that the *green* side wins the war of attrition increases over time. The *green* players thus face the problem of finding the optimal point in time at which to initiate change, where the trade-off is between the cost of further delaying change and the increased likelihood of prevailing in the war of attrition if change is initiated later.

Using a laboratory experiment with 9 treatments and 1080 participants we draw a comprehensive picture of the forces at work in the social change game. The experimental results reveal that 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 the literature has previously shown that 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, 2000)—we show that 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, we explore factors that can help groups escape the conformity trap. We find that exogenously reducing the cost of nonconformity promotes change but when we let participants choose the cost of nonconformity, they fail to lower them to a level at which change is likely to occur. We also find that about half of the participants underestimate how fast preferences change toward *green* and show that eliminating this bias helps promote change, as do opinion polls, rewarding pioneers of change and increased gains from social change. Finally, we examine the effects of group size and improved feedback.

Across all treatments, we find that the decisive factor leading to change is the presence of sufficiently many "principled" individuals who are willing to incur large material costs to instigate change. These individuals are committed to changing the norm, even though choosing *blue* throughout the game would earn them substantially higher payoffs. This finding is in line with research in sociology emphasizing the importance of a subset of highly interested and resourceful people who play a crucial role in the early phases of collective action (Oliver and Marwell, 2001; Centola et al., 2018).[5] We also find that

[3] Indeed, Oliver et al. (1985) write that "accelerating production functions underlie the mass actions popular associated with the term collective action, such as political demonstrations or revolutions. They are rare events relative to the grievances that might give rise to them, but they tend to accelerate once they start."

[4] Egorov and Harstad (2017) examine to what extent private activism can lead to corporate self-regulation and how this interacts with public regulation.

[5] Centola et al. (2018) report an experiment on tipping point dynamics and the emergence of new social conventions. Once a convention was established among all experimental participants, they introduced confederates who were directed to always choose the novel convention. Their findings show that when the size of the (exogenously imposed) committed minority reached 25% of the population, a tipping point was triggered. As in our study, the power of the initiators of change comes not from their authority or wealth but from their commitment to the cause. However, in our study, if

some of the individuals who are willing to lead change try to initiate change too early. To achieve change, individual attempts at initiating change must occur around the same point in time. Only then can a minority form which is large enough to tip the established norm. If some nonconformists move too early, when others haven't yet switched their preference, change will be harder to achieve. We also elicit risk and nonconformity preferences, which allows us to characterize initiators of change as individuals who have a greater tolerance for risk and stronger preferences for nonconformity. On the other hand, participants' gender doesn't correlate with the probability that an individual deviates from an established norm.

Other authors have relied on experiments to study the persistence of inefficient social institutions. 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. 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. This finding is replicated in Heggedal and Helland (2014). In a variant of the minimum-effort game of Van Huyck et al. (1990), 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. Brandts et al. (2014) study the effectiveness of leaders when inducing coordinated organizational change.[6] In concurrent research, Smerdon et al. (2016) and Duffy and Lafkyz (2018) both conceptualize norms as equilibria in a coordination game and investigate situations in which preferences over two alternatives change over time. Apart from these commonalities, the experimental settings differ from ours in terms of the matching technology, the way preferences change, the information structure, and the experimental treatments.[7] Despite the differences in design and focus, both studies document inefficient behaviors akin to our conformity trap.

The paper proceeds as follows. Section 2 introduces the social change game. Section 3 presents an analytical framework and a set of hypotheses. Section 4 presents the experimental design. Section 5 discusses the main results on the conformity trap, explores welfare implications and examines characteristics of instigators of change. Finally, Section 6 concludes.

2 The social change game

We consider an environment in which individuals' preferences change over time and interactions are governed by a social norm. There are n players. Preferences are given by a player's type $\theta \in \{B, G\}$. The benefit of a type θ player choosing $c \in \{b, g\}$ is denoted by $v_\theta(c)$. Type B players prefer *blue* (b) and type G players prefer *green* (g), i.e., we have $v_B \equiv v_B(b) - v_B(g) > 0$ and $v_G \equiv v_G(g) - v_G(b) > 0$.

change occurs, the group of committed individuals arises *endogenously*.

[6] See also Gërxhani and Bruggeman (2015) and Masiliūnas (2017) for experiments examining the ability of social groups to overcome coordination failures.

[7] In Smerdon et al. (2016) study participants don't know whether or in what way preferences change. As a result, many subjects erroneously believe that the majority of other group members prefers the status quo (pluralistic ignorance). Duffy and Lafkyz (2018) vary whether the change in preference is deterministic or stochastic. Interestingly, this manipulation has no significant effect on the occurrence of the conformity trap. In contrast, in our setting improving subjects' knowledge about how fast types change significantly increases the likelihood of the conformity trap.

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, players interact in periods $t = 1, \ldots, T$. In each period, they are matched into pairs at random and choose *blue* or *green.* 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 who chooses c_j is:

$$\begin{aligned}
\Pi_{\theta_i}^{bb} &= v_{\theta_i}(b) \\
\Pi_{\theta_i}^{gg} &= v_{\theta_i}(g) \\
\Pi_{\theta_i}^{bg} &= v_{\theta_i}(b) - n_g p \\
\Pi_{\theta_i}^{gb} &= v_{\theta_i}(g) - n_b p
\end{aligned} \tag{1}$$

where n_g is the number of other players choosing *green* and n_b the number of other players choosing *blue.* Throughout the paper we focus on situations in which $(n-1)p > v_\theta$ for $\theta = B, G$. That is, choosing the preferred color is never a dominant strategy.

A key feature of the game is that players' preferences over colors change over time. In period 1 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 γ. The rate of change γ is common knowledge. Players who have become type G never switch back to type B. That is, preferences gradually change from *blue* to *green.* The time horizon T is sufficiently long such that virtually everyone is expected to prefer *green* before period T is reached.

To illustrate the tension created in this environment, suppose that all players start by choosing *blue*, their preferred color in period 1. In subsequent periods, this creates a history of choosing *blue*, which serves as a focal point and potentially perpetuates the status quo to future periods. However, the potential for change to *green* increases over time because of the increasing number of type G players. An interesting question is therefore when and under what circumstances the emergence of leaders, who are willing to instigate change by deviating from the focal point, is most likely to occur, and whether or not the attempt at change will be successful.

3 Theoretical framework

The model presented in this section highlights the incentive trade-offs in the social change game. It is meant to motivate our experimental treatments and derive qualitative predictions.

The social change game incorporates two major trade-offs. The first one is related to the conflict between type B and type G players. We formalize this conflict as a war of attrition, adapting a recent model by Egorov and Harstad (2017). A novel feature of our model is that type G players choose when to start the war of attrition. In particular, they choose at what point in time to deviate from *blue* to *green*, thereby starting a phase of miscoordination. If type G players deviate from the *blue* norm early, most players are still type B and hence type G players are likely to lose the war of attrition. If they deviate late, type G players are numerous and more likely to prevail in the war of attrition, but the preferred color is chosen for a longer period of time before conflict is initiated.

The second trade-off arises within the group of type G players: Because the cost of change are highest for the individuals who deviate first, type G players have an incentive to wait for others to take the initiative. If everyone acts this way, change will not occur. This is akin to problems studied in the volunteering dilemma literature (e.g., Bliss and Nalebuff, 1984; Bilodeau and Slivinski, 1996; Bulow and Klemperer, 1999). Uncertainty about whether an instigator of change will be able to inspire enough others to follow further diminishes the willingness to lead change. We capture these challenges in reduced form as follows.[8] Let the *tipping point* be defined as the minimum number of players required to choose *green* such that a (myopic) type G player prefers to choose *green* as well. It is given by[9]

$$n^{TP} = \frac{n-1}{2} - \frac{v_G}{2p}. \tag{2}$$

The tipping point is the critical mass of players needed to switch to choosing *green* such that change becomes self-enforcing. We use it as a proxy for the difficulty of type G's coordination problem. Note that the same issue doesn't arise for type B players, as they are already coordinated on *blue*.

For the theoretical framework, we use a continuous time version of the social change game. Time runs from $t = 0$ to infinity with discount rate r. There are n players. Given the rate of change in preferences γ, the expected number of type B players after time t is $\mathrm{e}^{-\gamma t}n$. The expected number of type G players is $(1-\mathrm{e}^{-\gamma t})n$. We distinguish between two phases. In an initial phase, when everyone still chooses *blue*, type G players need to decide at which point in time t_1 they initiate an attempt at changing the norm to *green*. At this point, the game enters a war of attrition where both colors are chosen and players incur miscoordination costs. At each point in time during the war of attrition, both types decide whether to concede or continue the conflict. The status quo is retained if type G concedes first. Social change is observed if type B concedes first.

Players' optimal strategies in the war of attrition are determined by the cost incurred during the conflict and by the gains from succeeding in changing (or preserving) their preferred norm. The present value of winning the war of attrition is v_θ/r for type $\theta = B, G$.[10] The expected flow cost incurred during the miscoordination phase after duration τ are assumed to be

$$\begin{aligned} (1-\mathrm{e}^{-\gamma t_1})^2(n-1)p/y_B &\qquad \text{for type } B \text{ players} \\ \mathrm{e}^{-2\gamma t_1}(n-1)p/y_G &\qquad \text{for type } G \text{ players,} \end{aligned} \tag{3}$$

[8] To the best of our knowledge, there exists no model where one (or both) sides in a war of attrition faces a nested volunteering dilemma whose outcome determines the chances to win the war of attrition. While this seems to be a fruitful research avenue, developing such a model is beyond the scope of the present article.

[9] The tipping point n^{TP} solves $\frac{n^{TP}}{n-1}\Pi_G^{gg} + \left(1 - \frac{n^{TP}}{n-1}\right)\Pi_G^{gb} = \frac{n^{TP}}{n-1}\Pi_G^{bg} + \left(1 - \frac{n^{TP}}{n-1}\right)\Pi_G^{bb}$.

[10] We assume that players don't anticipate their own preference change. For the case when type B players do anticipate their own future preference change, their present value of winning the war of attrition equals $\frac{v_B}{r+\gamma} - \frac{v_G}{r+\gamma}\frac{\gamma}{r}$. Assuming that this present value is positive, the analysis remains qualitatively the same.

where

$$\begin{aligned} y_B =& \mu_B/\tau \\ y_G =& \mu_G/(\tau n^{TP}). \end{aligned} \tag{4}$$

For both types the numerator of (3) reflects the social norm. It equals the expected number of players of the opposite type—$(1 - e^{-\gamma t_1})(n - 1)$ for type B—times the penalty parameter p, multiplied by the fraction of players of the other type. The squared term arises because in the social change game miscoordination costs are incurred only when two matched players choose different colors and the probability to miscoordinate equals the fraction of players of the other type.

The expected miscoordination cost is then divided by y_B and y_G, which measure the respective types' ability to deal with conflict. The abilities to deal with conflict depend positively on the variables μ_B and μ_G. These variables are meant to capture risk-taking and nonconformity preferences for both types and are privately known to each type. Since it is difficult for individual players to predict these factors ex ante, we assume that a type θ player learns the realization of μ_θ only at t_1. In particular, at time t_1, μ_θ is independently drawn from an exponential distribution with expectation λ_θ. The probability density function is thus $f_\theta = \frac{1}{\lambda_\theta} e^{-\mu_\theta/\lambda_\theta}$.[11]

Further, y_G (but not y_B) depends negatively on the tipping point n^{TP}. As explained above, this reflects the idea that for large tipping points it will be more difficult for type G players to initiate and sustain an attempt at change. Finally, we assume that y_B and y_G decrease in the duration of conflict τ. The intuition is that as the welfare loss due to miscoordination becomes large, the desire to end the conflict increases. An alternative interpretation is to view the cost increase over time as a way of modeling a finite horizon. Although the horizon is infinite in our model, effectively, there exists a deadline after which continuing the war of attrition becomes too expensive.

The game described above has a unique Perfect Bayesian equilibrium. A detailed derivation can be found in Appendix A.

Proposition 1: *There is a unique Perfect Bayesian equilibrium characterized by a time t_1 at which type G players initiate social change (the war of attrition) and constant poisson rates ϕ_B and ϕ_G at which type B and type G concede, respectively.*[12]

Figure 1 depicts some predictions of the model, based on the parameters we will use in the experiment. In particular, it is shown how t_1 (the time when change is initiated), the probability that change is successful, and the duration of conflict depend on the returns to social change $v_G = v_G(g) - v_G(b)$, the penalty p and the group size n. It is important to note that when we vary n, we hold $(n-1)p$ constant, that is, we adjust p accordingly. This allows us to isolate the effect of a bigger group size keeping the

[11] We follow Egorov and Harstad (2017) in assuming that the values of the privately known variables are revealed only at the start of conflict and in using an exponential distribution. The main benefit of doing so is that it leads to a unique equilibrium. Note in particular that because μ_θ becomes private knowledge of type θ only at t_1, type G's choice of t_1 cannot signal information about μ_G.

[12] More precisely, given realizations μ_θ, each type θ plays a pure strategy, that is, each type chooses a time $t_1 + \tau_\theta(\mu_\theta)$ at which to concede the war of attrition if the other type doesn't concede before. However, since μ_θ is private information, from the other type's perspective, each type concedes at a constant poisson rate which depends on the expectation λ_θ of μ_θ.

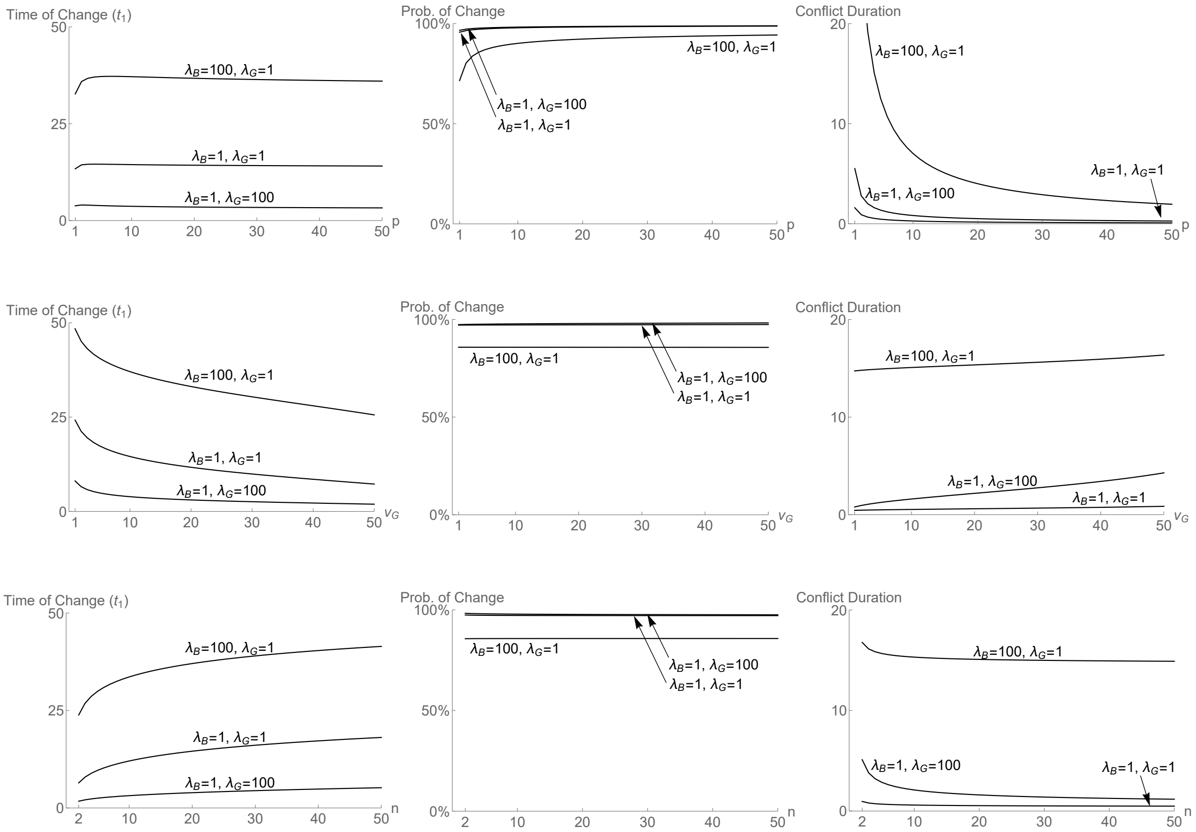


Figure 1: Model Predictions Based on Experimental Parameters

Notes: Theoretical predictions for the time of change initiation (left column), the probability of change (middle column), and the expected duration of the war of attrition (right column). Parameters fixed at $n = 20$, $p = 4$, $v_B(b) = v_G(g) = 30$ and $v_B(g) = v_G(b) = 20$ (unless varied on the x-axis). The rate of change in preferences equals $\gamma = Log(10/9)$ (continuous time equivalent of $\gamma = 0.1$ in the discrete model). Discount rate equals $r = 1/T$ where $T = 31$. Top row varies p, second row varies $v_G = v_G(g) - v_G(b)$, third row varies n holding constant $(n-1)p$. For each figure, different combinations of risk and nonconformity preference parameters λ_B and λ_G are shown.

expected cost of miscoordination constant. The figure also depicts outcomes for three combinations of players' expected risk-taking and nonconformity preferences (λ_B and λ_G).

The comparative statics are natural. Looking at the first column, we see that the initiation of change occurs earlier if type G players' gains from change are large (v_G), if groups are small (n), and if type G's risk-taking and nonconformity preferences (λ_G) are large relative to type B. Interestingly, the impact of the penalty parameter p on the timing of change initiation is negligible. This is because both sides incur miscoordination costs. For low penalties, both sides' equilibrium concession rates are low and hence conflict lasts a long time, and vice versa for high penalties. The main effect of p is thus on the duration of the war of attrition, see the third column in Figure 1. In the experiment, we will vary v_G, n and p, and we will elicit information about λ_θ.

The second column in Figure 1 reveals that the probability of change is almost invariant to changing

Both players are type B — Row: type G, Col: type B — Both players are type G

$$\begin{array}{c|c|c|}
 & b & g \\ \hline
b & \begin{array}{c}30\\30\end{array} & \begin{array}{c}30-n_g*4\\20-n_b*4\end{array} \\ \hline
g & \begin{array}{c}20-n_b*4\\30-n_g*4\end{array} & \begin{array}{c}20\\20\end{array} \\ \hline
\end{array}
\xrightarrow{\gamma=10\%}
\begin{array}{c|c|c|}
 & b & g \\ \hline
b & \begin{array}{c}20\\30\end{array} & \begin{array}{c}20-n_g*4\\20-n_b*4\end{array} \\ \hline
g & \begin{array}{c}30-n_b*4\\30-n_g*4\end{array} & \begin{array}{c}30\\20\end{array} \\ \hline
\end{array}
\xrightarrow{\gamma=10\%}
\begin{array}{c|c|c|}
 & b & g \\ \hline
b & \begin{array}{c}20\\20\end{array} & \begin{array}{c}20-n_g*4\\30-n_b*4\end{array} \\ \hline
g & \begin{array}{c}30-n_b*4\\20-n_g*4\end{array} & \begin{array}{c}30\\30\end{array} \\ \hline
\end{array}$$

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%

Figure 2: Social Change Game

Notes: Payoff matrices for parameters in Baseline treatment. The first entry in each cell is the row player's payoff, the second entry the column player's payoff. The variable n_c is the number of players choosing $c = \{b, g\}$. The lower part of the figure shows the expected percentage of type G players in each period for $\gamma = 10\%$.

p, v_G or n. The probability of change is high even if λ_B is much larger than λ_G. To understand why, note that the probability of change is predominantly determined by the discount rate r. If r is low, the probability of change is high because type G players don't lose much by waiting until they are an overwhelming majority. For instance, for low r, if v_G is lowered, type G react by waiting longer before initiating change in order to still reach a high probability of winning. In contrast, if r is high, type G players mainly care about the short term and initiate change early on, even if this means that they are less likely to prevail in the war of attrition. Note that $r = 1/T$. If we think about r as a breakdown probability, the expected number of periods equals $T = 31$, the fixed deadline used in the experiment.

4 Experimental design

This section describes the experiment. The experimental setting follows the description of the social change game in Section 2. In all treatments, we set the number of periods to $T = 31$ and the rate of change to $\gamma = 0.1$. The corresponding expected fraction of type G subjects in each period is given in the lower part of Figure 2. In expectation a majority prefers *green* by period 8 and the percentage of individuals preferring *green* reaches 90% by period 23.

4.1 Baseline

The Baseline treatment of the experiment is illustrated in the upper part of Figure 2. In each session, 20 subjects are randomly matched into pairs in each period and choose between *blue* (b) and *green* (g).

Table 1: Summary of Treatments

Treatment	Subjects[a]	Description (Difference to Baseline)
Establishing the Conformity Trap		
Baseline	120 (6)	$n = 20$, $T = 31$, $\gamma = 0.1$, $p = 4$, $v_B(b) = v_G(g) = 30$, $v_B(g) = v_G(b) = 20$.
Baseline+	120 (6)	Precise information about how many type G to expect in each period.
Low Penalty	120 (6)	Lower disunity penalty: $p = 1$.
Choose Penalty	120 (6)	Choose disunity penalty incurred by matched subject: $p = \{1, 4, 7\}$.
Gains from Change		
High Return	120 (6)	Higher payoff for type G if choosing g: $v_G(g) = 50$.
Reward	120 (6)	Initiators of change receive highest earnings when successful
Ability to Coordinate		
Small Group	60 (6)	Smaller group: $n = 10$. Moreover, $p = 8.44$ to keep $(n - 1) * p$ as in Baseline.
Feedback	120 (6)	Immediate feedback about everyone's color choice.
Poll	120 (6)	Poll asking about preferred color in period 14.

(a) Number of sessions/groups in parentheses.
Notes: All sessions were run at the economics laboratory of the University of California, San Diego between the fall of 2015 and 2016, except treatment Baseline+ in February 2019. The total number of participants is 1080 (including 60 subjects from which we elicited beliefs about the number of type G players not listed in the table above). Payments averaged \$36.1.

Initially, everyone prefers *blue*. In each period, each individual who has not yet switched to preferring *green* has a 10% probability that their preference switches to *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 penalty of 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 about the other subject'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.[13]

4.2 Escaping the conformity trap?

We implemented eight additional treatments to examine behavior in the social change game. All treatments are motivated by the idea to facilitate social change relative to the Baseline treatment. Table 1 presents an overview of the treatments. Due to the number of treatments, for clarity, we keep the description of the treatments brief in this section. For each treatment, more details will be provided when discussing the data.

Under the heading *Establishing the Conformity Trap*, we study the Baseline treatment, explore subjects' beliefs about the preference change (Baseline+) and examine the impact of the disunity penalty. The latter is a key variable as it determines the pressure to conform to others' behavior. We are interested in lowering the penalty (Low Penalty) as well as letting subjects choose the size of the disunity penalty (Choose Penalty). The Choose Penalty condition is operationalized by letting each subject in each period choose the disunity penalty of the subject they are matched to.

[13] Delaying feedback about aggregate group behavior is consistent with the pairwise interactions setting. Information about the choice of the other person in a match is instantly revealed, but behavior in the rest of the society is disseminated with a short delay. From a design perspective, we did this is to be able to study how the removal of the delay affects behavior (treatment Feedback).

Under the heading *Gains from Change*, we increase the benefits of social change for type G players in two ways. First, by increasing the value of choosing *green* as a type G from 30 to 50 (High Return), we introduce an asymmetry between types: type G have a larger incentive to follow their preferences than type B. Second, we incentivise the emergence of leaders by paying a reward to the first four subject who switch to choosing *green* and never go back to *blue* (conditional on change being successful). Outside the laboratory, rewards received by pioneers of change may take the form of privileged social positions, political power, fame, prizes, decorations, and monuments.

Under the heading *Ability to Coordinate*, we study different ways of improving coordination. In treatment Small Group, the group size is halved. Fewer subjects need to coordinate to reach the tipping point. In treatment Feedback, subjects immediately learn at the end of each period how many others in the group chose *blue* and *green* (recall that there is a one period delay in Baseline). Outside the laboratory, faster information dissemination may be a consequence of faster communication channels such as television and social media. Finally, we allow subjects to aggregate information and intentions by conducting a poll in period 14, asking which color they would prefer everyone to choose. The poll is anonymous. The outcome of the poll is announced to everyone.

4.3 Procedures

The experiment was conducted at the Economics Laboratory of the University of California, San Diego (UCSD) using the experimental software z-Tree (Fischbacher, 2007). A total of 1080 subjects participated in the experiment. We ran six sessions for each treatment. Each subject participated in one session only. 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, except treatment Baseline+ in February 2019. Participants were students at UCSD from various disciplines. The mean age was 20 years and 54% 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 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 \$36.1 per subject, including a show up fee of \$10. At the end of a session, we also elicited subjects' risk and nonconformity preferences.[14]

[14] In the risk task, subjects had to pick one of six lotteries: (a) 8 in 10 chance to win \$2, (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. See Andreoni and Harbaugh (2016). To elicit nonconformity preferences we asked subjects to rate statements taken from a scale measuring psychological reactance developed by Hong and Faedda (1996). A five-point rating scale from 1 ("strongly disagree") to 5 ("strongly agree") is used. The total score corresponds to the sum of the scores for the ten statements subjects had to consider. 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," "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.

4.4 Hypotheses

Notice that the socially efficient period of change, defined as the one that maximizes the ex-ante expected sum of payoffs, is given by the first period t^e at which $f_{G,t} \geq v_B/(v_B + v_G)$, where $f_{G,t}$ is the fraction of type G at time t.[15] For all treatments, $t^e = 8$ except for treatment High Return where $t^e = 4$ due to the larger value of v_G.

In the Baseline treatment, the cost for a type G player when individually deviating to *green* is 19*4-10=66 (19 other players choosing *blue* times the penalty of 4 minus the gain for choosing *green* instead of *blue*). This is a high cost compared to the long-term benefit of $v_G = 10$ per period if the deviation leads to change. We therefore expect that change occurs later than the socially efficient period of change (period 8), when in expectation only half of the players have switched preferences. Given this inefficiency, and following the model predictions discussed in Section 3, our main hypothesis is that each of the treatments increases the likelihood of change and/or leads to earlier social change compared to the Baseline treatment. A second hypothesis is that the elicited risk and nonconformity preferences positively correlate with the probability that a subject chooses to instigate social change.

5 Results

5.1 Establishing the conformity trap

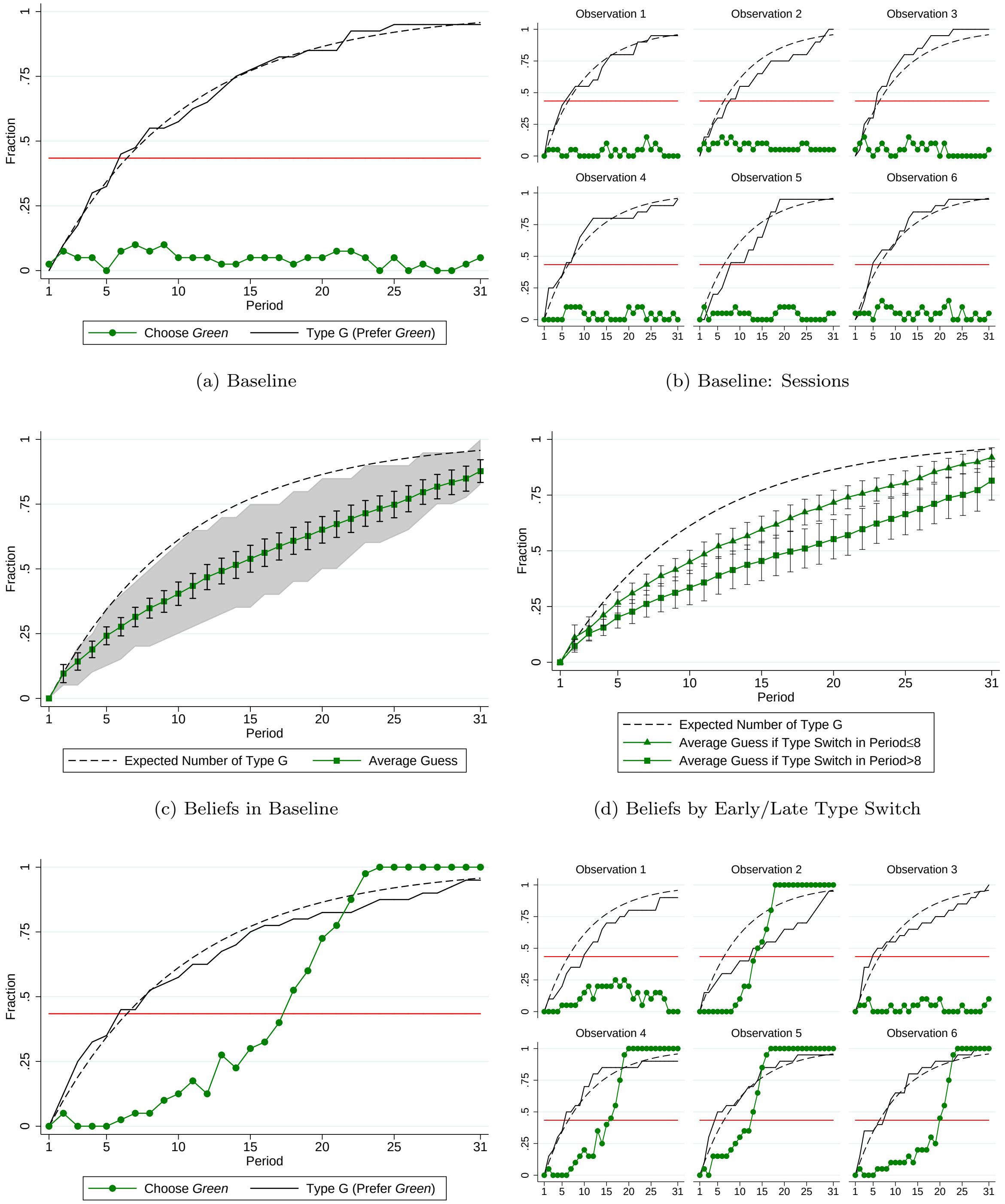
Figure 3 (a) and (b) present the results of the Baseline treatment. Figure (b) on the right-hand side shows the behavior in each session. Figure (a) summarizes behavior by depicting the median observation over the six sessions. The green line with circled markers depicts the fraction of players choosing *green*. The solid increasing line shows the realized fraction of type G and the dashed line shows the theoretical expectation. Finally, the horizontal line depicts the tipping point.

Result 1 (Baseline)**:** *All groups are caught in the conformity trap. That is, blue is chosen by a large majority throughout the game, even when almost all subjects prefer green.*

Support: Figure 3b shows that in all six sessions of treatment Baseline, groups chose *blue* up to and including period 31. Color *blue* was chosen 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. □

Result 1 raises a number of questions. First, is it surprising that none of the six group achieved change? On the one hand, in line with our hypothesis, some delay before change occurred should be expected. On the other hand, the lock-in at *blue* occurs despite the fact that the change in preferences over time is common knowledge. Figure 3b shows that in each session there were many periods with isolated attempts at instigating change, leading to up to three simultaneous *green* choices. Hence, the miscoordination cost didn't prevent deviations but the deviations weren't sufficiently frequent or

[15] The expression is derived from $f_{G,t} n v_G = (1 - f_{G,t}) n v_B$. Given a sufficiently large miscoordination penalty p, in the efficient equilibrium, all players choose the same color in all periods (at time t^e, all players switch from *blue* to *green*).



(a) Baseline

(b) Baseline: Sessions

(c) Beliefs in Baseline

(d) Beliefs by Early/Late Type Switch

(e) Baseline+

(f) Baseline+: Sessions

Figure 3: Baseline and Baseline+

Notes: Circled markers show the fraction of subjects choosing *green*. The solid increasing line shows the fraction of type G subjects; the dashed line the corresponding expectation. The horizontal line is the tipping point. Figures (a) and (e) depict the median observation over the six sessions shown in Figures (b) and (f). Figure (c) shows the average guess about the number of type G subjects from the belief elicitation sessions and the 95% confidence intervals. Shaded area extends from 25th to the 75th percentile. Figure (d) separates by whether a subject became type G before or after period 8.

coordinated to reach the tipping point. We will study the characteristics of subjects who tried to instigate change in Section 5.5.

Second, to what extent does the lock-in at *blue* depend on the size of the penalty? In particular, if subjects can choose the penalty themselves, would we still observe the conformity trap? We address this question in treatments Low Penalty and Choose Penalty below.

Third, one may wonder how good subjects were at computing the change in preferences (change in types). While everyone knows that the rate of change is 10%, subjects may miscalculate and not be aware that this implies that after some period a large majority prefers *green*, with a probability close to 100%. To address this question, we ran 3 additional sessions in which we elicited subjects' beliefs. Subjects went through the 31 periods where, just as in the Baseline treatment, subjects knew their own type and are informed about the rate of change of 10%. However, instead of playing the social change game, in each round they had to guess the number of type B and G participants. Subjects were paid according to the accuracy of their guesses.[16] Figures 3 (c) and (d) show the results for the 60 subjects participating in these sessions. The left figure shows that on average subjects guessed significantly fewer type G players than there actually were. The right figure shows that the bias exists in both groups of subjects, those who switched type early—before period 8, the point in time where ex ante each subject has a 50% probability of having changed type—and those who switched type late, although the bias is larger for the ones who experience late type changes. It is worth noting that 40% of the subjects make guesses that are fully in line or close to the theoretical expectation. However, the ones who don't guess correctly underestimate the rate of change.

The bias in beliefs about how fast types change suggests that establishing common knowledge about the expected speed of the preference change may help groups escape the conformity trap. To test this, we ran treatment Baseline+. The treatment is identical to the Baseline treatment except that in the instructions subjects are presented with a table showing how many of the 20 participants were type G in each round of the six previously conducted Baseline sessions. Figures 3 (c) and (d) present the results for Baseline+.

Result 2 (Baseline+)**:** *Establishing common knowledge about the expected number of type G players over the 31 periods facilitates change to green.*

Support: Figure 3 (d) shows that in four out of six sessions of Baseline+ groups achieved change to *green*. This is a significant difference relative to the Baseline treatment (one-sided Fisher exact test, p=0.030), even though two groups were still locked in at *blue*. It is also worth noting that when change occurs in Baseline+, the tipping point is reached around period 15, later than the socially efficient period of change $t^e = 8$. □

Baseline+ is a strong intervention. In addition to correcting individual biases in beliefs, the treatment also establishes common knowledge that everyone has correct beliefs. Furthermore, by showing subjects how types changed over time in the six Baseline sessions, they learn that the variance between the sessions is small (due to the large groups, realized type changes are usually close to the expected value).

[16] At the end, one of the rounds was randomly selected. If the difference between a subject's guess and the true number of type G participants in the selected round was 0, 1, 2, 3, 4, 5 or 6 the subjects earned \$36, \$35, \$33, \$29, \$24, \$18 or \$11, respectively. For higher differences, subjects earned \$5.

These informational conditions are unlikely to be satisfied in most real world contexts.

Also note that it takes considerable persistence on the part of the first deviators to *green* to convince others to change their choice as well. After the occurrence of the initial deviation to *green* that eventually lead to change, it took at least 10 periods until the tipping point was reached (with the exception of session 2). Initiators of change therefore incur large losses.

We now turn the discussion to the treatments varying the penalty parameter. In Low Penalty, the cost to nonconformity is exogenously reduced to $p = 1$. All other aspects of the environment are identical to the Baseline treatment.

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

Support: Figures 4 (a) and (b) show that lowering the disunity penalty parameter to $p = 1$ almost eliminates the lock-ins at the *blue* norm. 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, as the switch to *green* still occurs after period 8 (change did not occur in session 1) and the transition process still takes more than 10 periods.[17] □

In many of the examples discussed in the introduction individuals have some control over the extent to which they want to sanction others who violate a norm. Addressing the robustness of the conformity trap, an interesting question is thus whether subjects choose low or high penalties, if they have a choice. In treatment Choose Penalty, in each round, we let subjects choose the miscoordination penalty incurred by the other subject in the pairwise match. Subjects can choose between a low penalty ($p = 1$ as in Low Penalty), a medium penalty ($p = 4$ as in Baseline) or a high penalty ($p = 7$).[18] Figures 4 (c) and (d) show the results for treatment Choose Penalty.

Result 4.1 (Choose Penalty)**:** *The likelihood of groups getting caught in the conformity trap in Choose Penalty is not lower than in Baseline.*

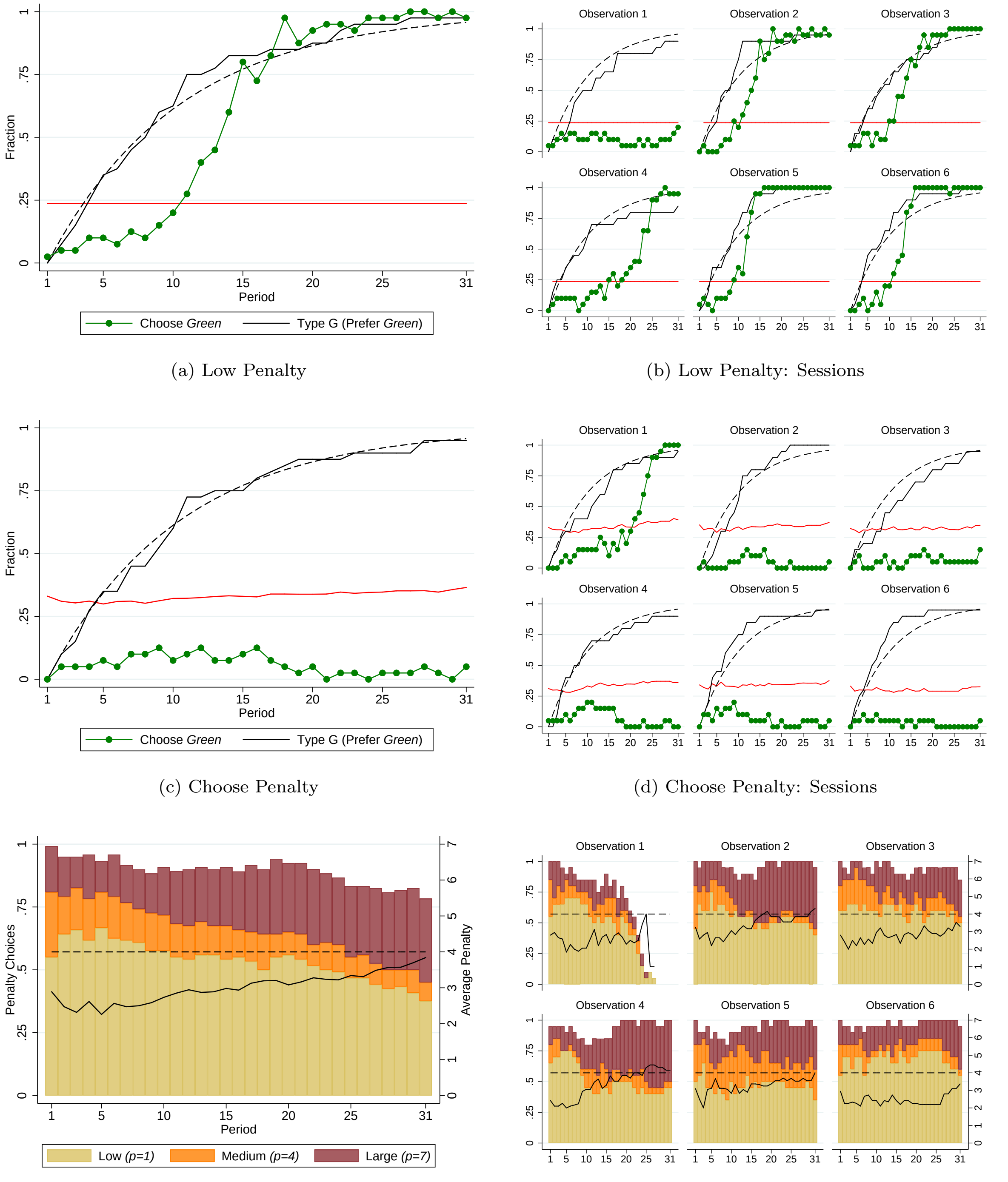
Support: Five out of six groups in Choose Penalty were unable to escape the conformity trap, see Figure 4 (d). Social change is not significantly more likely than in the Baseline treatment (one-sided Fisher's exact test, p=0.500), despite the option to choose $p = 1$, which corresponds to the value in Low Penalty in which change was frequently observed. □

To understand why the possibility of choosing penalties is not sufficient to avoid the conformity trap, we take a closer look at subjects' penalty choices. Figures 4 (e) and (f) depict 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, that is, the average penalty faced by the players who choose *green*.

Result 4.2 (Penalty Choices in Choose Penalty)**:** *Penalty choices of subjects choosing blue (i.e., penalties against green) increase over time. This is true for type B as well as type G, suggesting that medium*

[17] The latter finding is expected. As implied by our model, lower penalties tend to lead to longer phases of miscoordination since type G and type B will both face a lower cost when sticking to their preferred color.

[18] We allow for a penalty of 7 so that type B players can potentially counteract low penalty choices of type G players.



(a) Low Penalty

(b) Low Penalty: Sessions

(c) Choose Penalty

(d) Choose Penalty: Sessions

(e) Penalty Choices of Subjects Choosing *blue*

(f) Penalty Choices: Sessions

Figure 4: Low Penalty and Choose Penalty

Notes: Circled markers show the fraction of subjects choosing *green*. The solid increasing line shows the fraction of type G subjects; the dashed line the corresponding expectation. The horizontal line is the tipping point. Figures (e) and (f) show the fraction of low ($p = 1$), medium ($p = 4$) and large ($p = 7$) penalty by subjects choosing *blue* (penalties against *green*) and the corresponding average penalty. The total height of each bar corresponds to the fraction of subjects choosing *blue*. Figures (a), (c) and (e) depict the median observation over the six sessions shown in Figures (b), (d) and (f).

and high penalties are chosen out of a fear of miscoordination.

Support: Figures 4 (e) and (f) show that the average penalty choice is below $p = 4$ in most periods. However, there is a significant upward trend in average penalties (Spearman's rho above 0.7). This is true irrespective of subjects' types, i.e., it is true even for subjects who prefer *green*: 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, $p = 0.66$). □

The behavior observed in the Low Penalty and Choose Penalty conditions highlights an interesting trade-off: If disunity penalties are low, beneficial change is more likely to occur, but due to the uncertainty about whether change will in fact be realized and how long the transition will last, subjects also have an incentive to choose high penalties to avoid costly miscoordination. This concludes the discussion of the first part of the results section. We have documented the existence of conformity traps and shown that they are exacerbated by biased beliefs about others' preferences and a fear of miscoordination.

5.2 Gains from change

In this section, we discuss two treatments varying the returns to change. In the first case, all type G players have more to gain when achieving change. In the second case, an endogenously determined subset of individuals receives a large reward for initiating change, but incentives remain as in the Baseline for the rest of the group members.

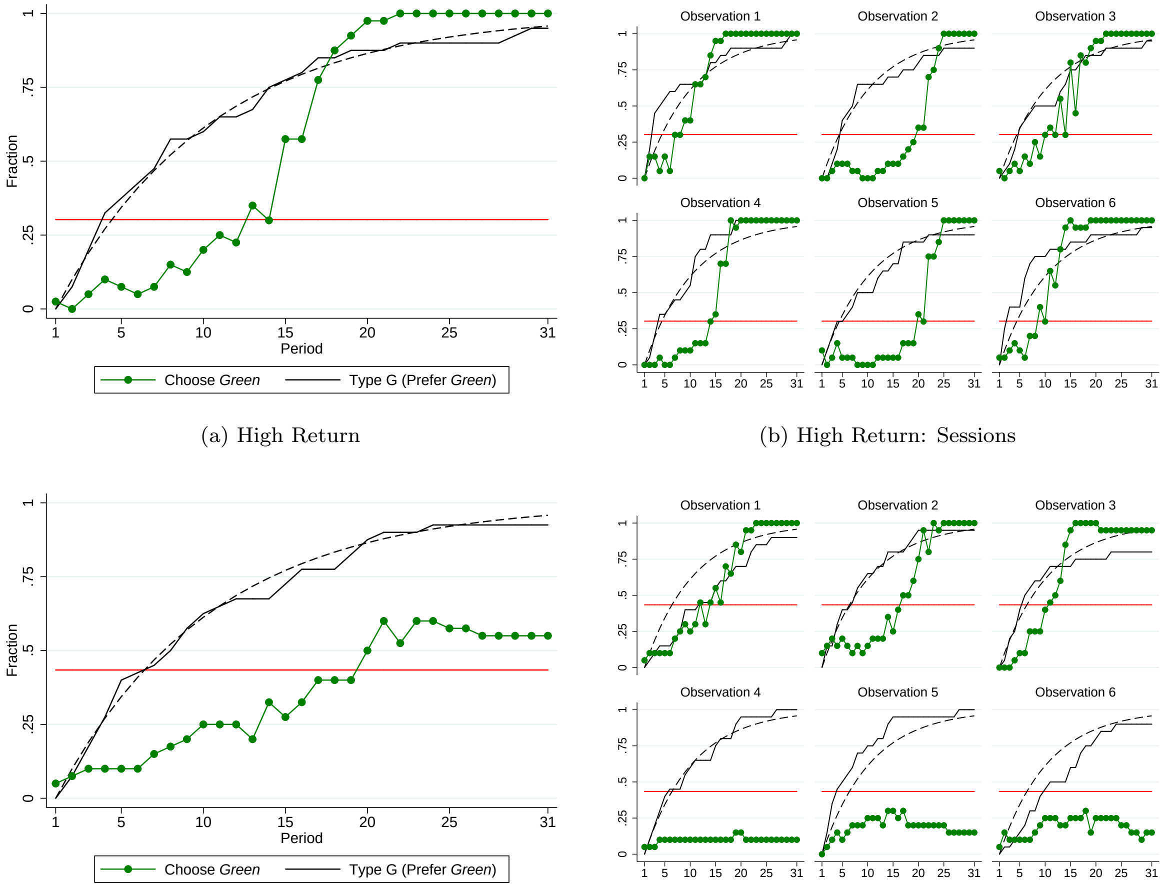
Treatment High Return provides a test of the prediction that increasing the value of choosing *green* for type G from $v_G(g) = 30$ to $v_G(g) = 50$ (increasing v_G from 10 to 30) leads to earlier change. Figures 5 (a) and (b) depict the results.

Result 5 (High Return)**:** *Higher returns for type G subjects when choosing green facilitate change, although change still occurs significantly later than in the efficient outcome.*

Support: In all six sessions groups broke out of the conformity trap, a significant difference to the Baseline treatment (one-sided Fisher's exact test, p=0.001). 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 *green* should occur in period 4 (earlier than in the other treatments). □

Notice that change tends to be slow until the tipping point is reached and speeds up thereafter. This suggests that change primarily depends on a subset of individuals who are willing to make the first steps toward change. Treatment Reward incentivises the emergence of such individuals. Define the majority color as the color chosen by more than 50% of the 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*.[19] Initiating change to *green* thus promises a reward, but it is also risky in case change fails to occur. The reward of the "top four" subjects is that their earnings are raised to the level of the highest-earning subject in the session. Figures 5 (c) and (d) show behavior in treatment Reward.

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



(a) High Return

(b) High Return: Sessions

(c) Reward

(d) Reward: Sessions

Figure 5: High Return and Reward

Notes: Circled markers represent the fraction of subjects choosing *green*. 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. Figures (a) and (c) depict the median observation over the six sessions shown in Figures (b) and (d).

Result 6 (Reward)**:** *Rewards for initiators of change help groups adapt to green in half of the sessions.*

Support: Three groups were caught in the conformity trap and three groups escaped it. The difference to Baseline is significant at the 10% level (one-sided Fisher's exact test, p=0.090). □

Notice that in sessions 4, 5 and 6 several subjects chose *green* early on, trying to earn the Reward, but eventually failed to induce change. These subjects incurred large losses and were also harmful for the group because of the miscoordination costs they caused. This is illustrated by the low average per period 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). Further, recall that we elicited subjects' risk preferences at the end of each session.[20] Not surprisingly, the subjects who deviated first to receive the reward tend to tolerate risk better. Because the less risk-averse individuals

[20] An analysis of the risk elicitation measure can be found in Section 5.5.

are also the ones who are most likely to choose *green*, even in the absence of a reward, the reward would be more effective if paid to the "marginal" subject required to reach the tipping point (as opposed to the initial deviators). A key conclusion is thus that in the type of setting we examine incentivizing followers to join a movement is as critical to beneficial change than promoting the emergence of leaders.

5.3 Ability to coordinate

This section examines our last set of treatments, designed to promote change by helping groups better coordinate the break out of the conformity trap. We manipulate the group size, improve the feedback about others' behavior, and give participants the opportunity to aggregate preferences via a public opinion poll.

Our model predicts that change is more easily achieved in smaller groups, since less coordination is required to reach the tipping point. In treatment Small Group, the group size is reduced from $n = 20$ to $n = 10$. Simultaneously, we increase the disunity penalty parameter from $p = 4$ to $p = 8.44$ in order to keep the expected cost of miscoordination the same as in the Baseline treatment. 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. Figures 6 (a) and (b) depict the results for Small Group.

Result 7 (Small Group)**:** *Smaller groups are more likely to escape the conformity trap but suffer from large disunity costs due to long transition phases from blue to green.*

Support: Three groups were caught in the conformity trap and three groups escaped it, a (weakly) significant difference to the Baseline treatment (one-sided Fisher's exact test, p=0.090). In the sessions where change was achieved, the transition from *blue* to *green* was slow and miscoordination costs large. 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 per period payoff of 19.79 in the sessions without change. □

In Section 5.4, we will show that treatment Small Group was the least efficient among all treatments, mainly due to the large miscoordination costs. In smaller groups, conflict between individuals preferring *green* and *blue* seems to be particularly strong. A possible explanation is that subjects are aware that their choice has a bigger impact on the group norm and are thus more motivated to endure miscoordination.

In treatment Feedback, subjects receive immediate feedback about the number of people who chose *blue* and *green* at the end of each period. Compared to treatment Baseline, where this information is delayed by one period, we expect this to improve subjects' ability to signal a willingness to change and hence their ability to coordinate. Figures 6 (c) and (d) summarize the behavior observed in treatment Feedback.

Result 8 (Feedback)**:** *Expediting feedback does not reduce the likelihood of groups getting caught in the conformity trap compared to treatment Baseline. In fact, deviations to green are less common in treatment Feedback than in treatment Baseline.*

Support: Change to *green* occurred in only one out of six sessions, not significantly different from the Baseline treatment (one-sided Fisher's exact test, p=0.500). However, providing immediate feedback about the color choices in the group did affect behavior. Subjects in treatment Feedback were on average *less* likely to choose *green* than participants in the Baseline treatment, as is apparent from the fact that the line depicting the fraction of subjects choosing *green* is flatter than in Baseline. In Section 5.5, we will present regressions corroborating this finding and its statistical significance. □

Faster feedback generally reduces the probability of deviations to *green*, but, at the same time, the late but rapid change to *green* observed in session 3 suggests that faster feedback can also speed up change once it has been initiated. In other words, increasing the speed of feedback seems to lead to more extreme behavior, that is, for a long time deviations from a norm may be rare but once the norm is challenged, rapid change can happen. This is a preliminary observation and we hope future studies will provide conclusive evidence on the impact of delaying and speeding up feedback in the social change game.

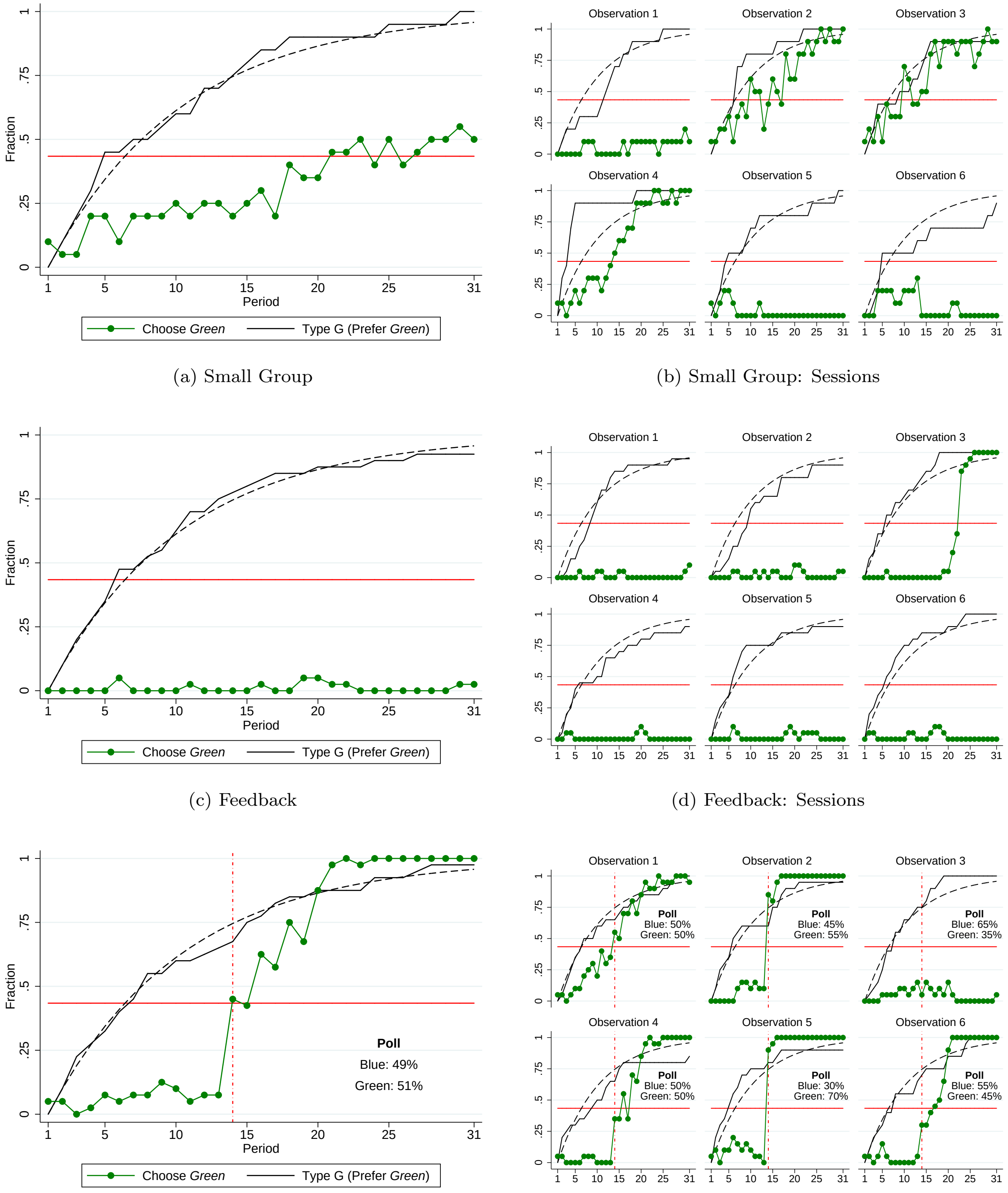
We conclude with treatment Poll. At the start of period 14, subjects are asked what color they would prefer people in their matching group chose in the next rounds. After learning how many people answered *blue* and how many *green*, all subjects make their actual color choices for period 14. All aspects of the poll are explained in the experiment instructions and subjects are aware at the start of the game that there will be a poll. In period 14, in expectation 75% of the subjects prefer *green* and the probability that the group will have a majority preferring *green* is 98.5%. The poll should therefore function as an effective coordination device.

The results for Poll are presented in Figures 6 (e) and (f). The figure includes as additional information the percentage of subjects who stated *blue* or *green* as their preferred color in the poll.

Result 9 (Poll)**:** *The poll enables groups to break out of the conformity trap, even though a significant fraction of type G players vote for blue.*

Support: Polls are effective coordination devices in our experiment, leading to change in five out of six sessions, a significant difference to the Baseline treatment (one-sided Fisher's exact test, p=0.008). The group in session 3 of the Poll treatment did not manage to escape the conformity trap. In this session, 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 register a preference for *green* in the poll 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. □

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. 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 vote for *blue* despite a private preference for *green* and despite



(a) Small Group

(b) Small Group: Sessions

(c) Feedback

(d) Feedback: Sessions

(e) Poll

(f) Poll: Sessions

Figure 6: Small Group, Feedback and Poll

Notes: Circled markers represent the fraction of subjects choosing *green*. 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. 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.

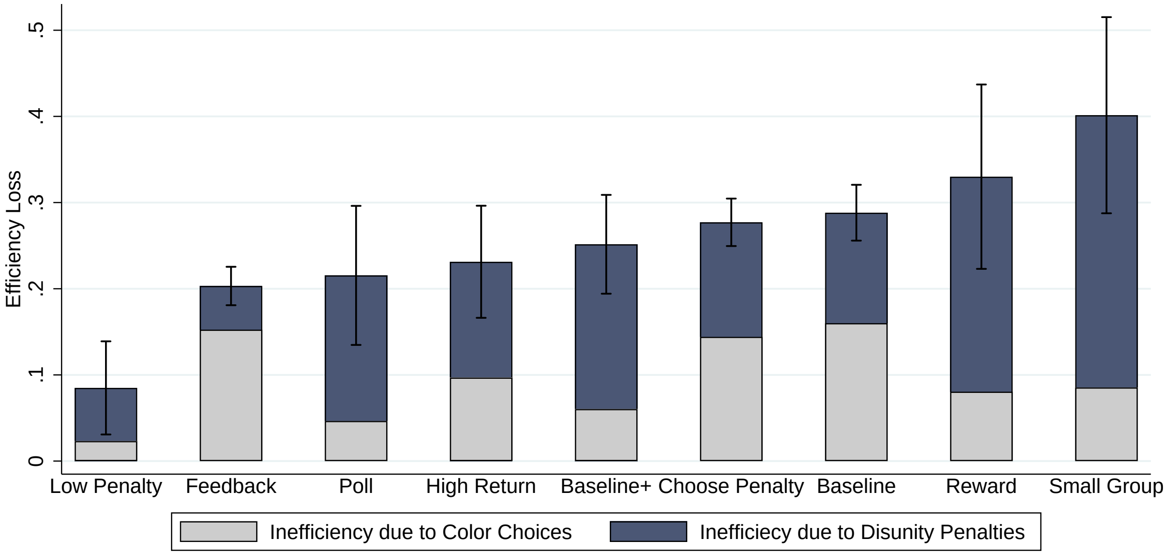


Figure 7: Efficiency Loss Relative to Social Optimum

Notes: Percentage of payoff loss relative to the efficient outcome with 90% confidence intervals. The lower part of each bar shows the inefficiency due to color choices; the upper part the inefficiency due to disunity penalties.

the fact that the poll is anonymous.[21] Notice that a similar fear of miscoordination was already visible in treatment Choose Penalty, where type G players opted for high penalties to discourage attempts at change.

5.4 Efficiency

Results 1 to 9 document that groups often fail to adapt to the *green* norm. What are the implications for social welfare? In Figure 7, 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 10 (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 7 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, 36% of the efficiency loss is due to delayed or no social change (i.e., not choosing the preferred color), 64% is due to miscoordination (i.e., disunity penalties). 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). Feedback

[21] Interestingly, if we go back to Figure 3 (e), the elicited beliefs show that on average people believed that 50% are type G in period 14, suggesting that without the bias in beliefs the poll would have lead to quicker change.

is more efficient than Baseline (p=0.004). The least efficient treatment is Small Group (p=0.078 if compared to Baseline). The differences in efficiency between Baseline and the remaining conditions is insignificant based on Mann-Whitney U tests.[22] □

5.5 Initiators of change

An interesting feature in our data is that the build-up of *green* choices toward reaching the tipping point tends to be slow. The ability of groups to achieve change thus critically depends on individuals who are willing to persistently choose *green* when most others still choose *blue*. We call them *initiators of change* or *nonconformists*.

Figure 8 (a) focuses on nonconformists' initial deviations to *green*. For each subject who chose *green* at least once when the tipping point has not been reached yet, we plot the period when the subject chose *green* for the first time (y-axis) against the period in which this subject's type changed from type B to G. Only few choices lie below the 45 degree line, indicating that type B players rarely tried to initiate change. More interestingly, many nonconformists deviated to *green* in the same period as their type changed or shortly thereafter. If such attempts occur early, say before period 10, the chance that they cause change is small. As predicted in our model, a more promising strategy for type G players would have been to concentrate deviations to *green* on a period where sufficiently many participants have already changed type. This is easier said than done. However, the point is that many nonconformists have an impulse to deviate to *green* as soon as their preferences change, which, perhaps counter-intuitively, diminishes their chances of triggering change.[23]

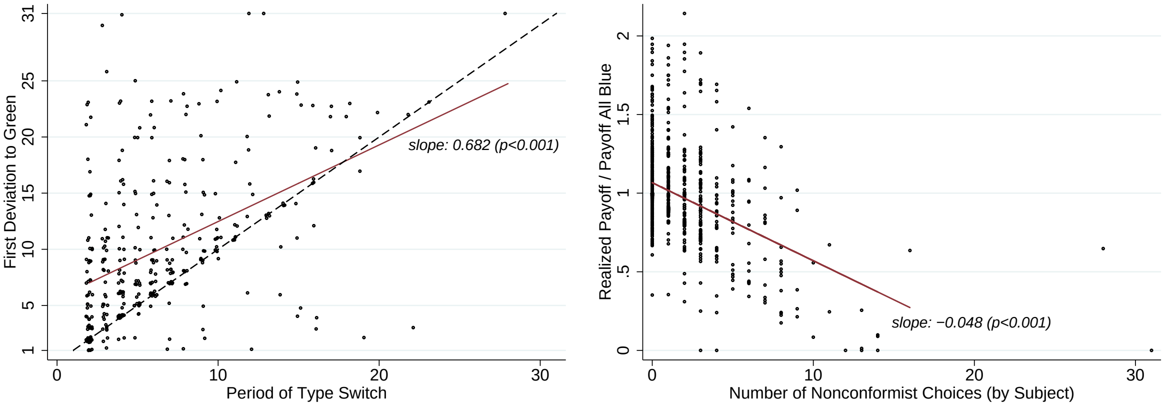
Figure 8 (b) illustrates the cost of initiating change. The figure shows for each subject their realized payoff divided by the payoff they would have received if everyone in the group chose *blue* in all periods (y-axis). This is plotted against the number of times a subject chose *green* when the tipping point has not been reached yet (x-axis). We see that this ratio is above 1 for many of the subjects who never or rarely make a nonconformist choice: These subjects benefitted from the nonconformists' efforts to overcome the status quo. On the other hand, subjects with multiple nonconformist choices are almost always worse off than they would be if no attempts at change were ever made. On average, each nonconformist choice reduces earnings by 4.8% relative to the all *blue* outcome. Despite the high cost, many subjects are persistent nonconformists, often losing more than 50% compared with the "all blue" outcome. Hence, the actions of initiators of change can only be explained by non-pecuniary motives.[24]

To better understand these motives, Table 2 provides regressions with the dependent variable being the

[22] Notice that most observations in Poll are substantially more efficient than each of the six sessions in Baseline (the overall p-value of 0.150 is due to Session 1 and 3, which are slightly less efficient than the average Baseline session).

[23] The difficulty to coordinate is also illustrated by the fact that, across all treatments, if change wasn't achieved, there were on average 2.5 times as many subjects who chose *green* at least once than the maximum number of simultaneous *green* choices. For instance, in treatment Baseline, where the maximum number of simultaneous *green* choices was three in most sessions, this information tells us that on average there were 7.5 nonconformists. If all nonconformists deviated around the same period, change would likely have happened. If change was achieved, coordination was significantly better: On average, it only took 1.6 nonconformists to increase the maximum number of *green* choices by 1.

[24] In the online appendix, the same analysis is done except that realized payoffs are normalized by the payoff if best-responding to the other participants' actual behavior. The cost of leading change remains at about 5% per nonconformist choice.



(a) Timing of Deviations

(b) Cost of Leading Change

Figure 8: Initiators of Change

Notes: Figure (a): Time of first deviation to *green* plotted against the period in which the subject switched from type B to type G. Includes only deviations occurring when the tipping point has not been reached yet. Change means that more than 50% of the subjects choose *green* in the last period. Figure (b): Realized payoff divided by hypothetical payoff if everyone chose *blue* in all periods plotted against the number of nonconformist choices (choosing *green* when the tipping point has not been reached yet). Data include all treatments except Reward.

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). It equals 0 for subjects who chose lotteries that imply some degree of risk-aversion and 1 otherwise. We also elicited subjects' nonconformity preferences (see Section 5.2). If the score of a subject in the survey exceeded the median score among the 1080 participants, we classify her as a nonconformist.[25]

The first insight we can derive from these regressions is that all treatments except treatment Feedback and treatment Poll have a significant positive effect on the probability that subjects try to instigate change by deviating to *green* (relative to the Baseline treatment). For instance, while we previously saw that treatment Choose Penalty didn't help groups overcome the conformity trap, the regressions show that there were more attempts at change than in the Baseline treatment. The fact that the regressions show an insignificant effect for treatment Poll is in line with the idea of the treatment: No isolated attempts at change are needed if groups can coordinate on initiating change in period 14 when the poll was conducted. Finally, the significant negative effect of treatment Feedback confirms the previous observation that there are less deviations when feedback is more immediate.

The key message of Table 2 is the significant effect of both the risk and the nonconformity measure on subjects' probability of initiating change.

Result 11 (Risk and Nonconformity)**:** *Initiators of change tend to have a greater tolerance for risk and a greater dislike for conformity.*

[25] 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 looked at political ideology and ethnicity and found these variables to be insignificant and not significantly affecting any of the other coefficients. The results are also robust to using the tipping point as the fraction of *green* choices below which an observation is included (see the online appendix).

Table 2: Determinants of the Probability of Choosing *Green*

Prob($g|\theta = G$)	(1)	(2)	(3)	(4)
Baseline+	0.030*** (0.007)	0.031*** (0.007)	0.029*** (0.007)	0.030*** (0.007)
Low Penalty	0.048*** (0.006)	0.047*** (0.006)	0.048*** (0.006)	0.048*** (0.006)
Choose Penalty	0.016** (0.007)	0.016** (0.007)	0.016** (0.007)	0.015** (0.007)
High Return	0.029*** (0.006)	0.029*** (0.006)	0.029*** (0.006)	0.029*** (0.006)
Reward	0.027*** (0.009)	0.027*** (0.009)	0.027*** (0.009)	0.027*** (0.009)
Small Group	0.028** (0.011)	0.027** (0.010)	0.027** (0.011)	0.027** (0.011)
Feedback	-0.021*** (0.005)	-0.021*** (0.004)	-0.022*** (0.004)	-0.022*** (0.004)
Poll	0.013 (0.009)	0.012 (0.009)	0.012 (0.009)	0.012 (0.010)
Risk-Accepting	0.018*** (0.005)		0.017*** (0.005)	0.017*** (0.005)
Nonconformist		0.012** (0.004)	0.011** (0.004)	0.011** (0.005)
Female				-0.003 (0.005)
Period Dummies	Yes	Yes	Yes	Yes
Observations (Individuals)	12,516 (819)	12,516 (819)	12,516 (819)	12,516 (819)
Clusters (Sessions)	54	54	54	54

Notes: * 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 the number of subjects choosing *green* is below 20% (i.e. 2 or less in Small Group and 4 or less in all other treatments). The reference treatment is Baseline.

Support: Table 2 shows that risk-accepting subjects were more likely to be initiators of change. Their probability of choosing *green* is 1.7 to 1.8 percentage points higher than the probability of choosing *green* of risk-averse subjects. Notice that this is differential is substantial because it applies in each period and hence accumulates over time. The coefficient of the non-conformity measure is also significant, indicating that initiators of change derive utility from leading change offsetting the material costs of doing so. □

6 Concluding remarks

Our study provides evidence that social norms 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 and when the pressure to conform is endogenous. 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.

Facilitating change 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 large losses in the short run in order to lower the cost for the next group, and so on until a cascade of change comes. We show that risk-accepting individuals as well as individuals who score high in a nonconformity survey are most likely to lead change. A second 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 that people tend to underestimate the speed of the preference change. Establishing precise common knowledge about the evolution of preferences helps promote change. Fourth, in several treatments we observe a fear of miscoordination which leads to conservatism. For example, the fear of miscoordination reduces the effectiveness of public opinion polls, because some individuals hide their true preference, even when the poll is anonymous.

We view our study as a starting point to a research program investigating empirically the impact different forces have on social change in a controlled environment. In most instances, assessing the efficiency of social change based on naturally occurring data is problematic due to the paucity of reliable data on individual preferences in daily life and the difficulty of aggregating them. Our experimental paradigm can be easily extended to explore richer environments, for instance when individuals interact in social networks.

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.

Akerlof, George, "The economics of caste and of the rat race and other woeful tales," *The Quarterly Journal of Economics*, 1976, pp. 599–617.

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

Andreoni, James and B Douglas Bernheim, "Social image and the 50–50 norm: A theoretical and experimental analysis of audience effects," *Econometrica*, 2009, *77* (5), 1607–1636.

_ **, William Harbaugh, and Lise Vesterlund**, "The Carrot or the Stick: Rewards, Punishments, and Cooperation," *The American Economic Review*, 2003, *93* (3), 893–902.

Arrow, Kenneth J., "Political and economic evaluation of social effects and externalities," *Frontiers of Quantitative Economics*, 1971, pp. 13–25.

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

Asch, Solomon E., "Studies of independence and conformity: I. A minority of one against a unanimous majority," *Psychological monographs: General and applied*, 1956, *70* (9), 1.

Balafoutas, Loukas, Nikos Nikiforakis, and Bettina Rockenbach, "Direct and indirect punishment among strangers in the field," *Proceedings of the National Academy of Sciences*, 2014, *111* (45), 15924–15927.

Bernheim, Douglas B., "A theory of conformity," *The Journal of Political Economy*, 1994, *102* (5), 841–877.

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

_ , *Norms in the wild: How to diagnose, measure, and change social norms*, Oxford University Press, 2017.

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

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch, "A theory of fads, fashion, custom, and cultural change as informational cascades," *The Journal of Political Economy*, 1992, *100* (5), 992–1026.

Bilodeau, Marc and Al Slivinski, "Toilet cleaning and department chairing: Volunteering a public service," *Journal of Public Economics*, 1996, *59* (2), 299–308.

Bliss, Christopher and Barry Nalebuff, "Dragon-slaying and ballroom dancing: The private supply of a public good," *Journal of Public Economics*, 1984, *25* (1-2), 1–12.

Boehm, Christopher, *Blood revenge: The anthropology of feuding in Montenegro and other tribal societies*, University Press of Kansas, 1984.

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.

_ , David J Cooper, and Roberto A Weber, "Legitimacy, communication, and leadership in the turnaround game," *Management Science*, 2014, *61* (11), 2627–2645.

Bulow, Jeremy and Paul Klemperer, "The generalized war of attrition," *American Economic Review*, 1999, *89* (1), 175–189.

Bursztyn, Leonardo, Alessandra González, and David Yanagizawa-Drott, "Misperceived social norms: Female labor force participation in Saudi Arabia," Technical Report, National Bureau of Economic Research 2018.

_ , Georgy Egorov, and Stefano Fiorin, "From extreme to mainstream: How social norms unravel," Technical Report, National Bureau of Economic Research 2017.

Centola, Damon, Joshua Becker, Devon Brackbill, and Andrea Baronchelli, "Experimental evidence for tipping points in social convention," *Science*, 2018, *360* (6393), 1116–1119.

Cinyabuguma, Matthias, Talbot Page, and Louis Putterman, "Cooperation under the threat of expulsion in a public goods experiment," *Journal of Public Economics*, 2005, *89* (8), 1421–1435.

Coleman, James S., "The emergence of norms," *Social institutions: Their emergence, maintenance, and effects*, 1990, pp. 35–39.

_ , *Foundations of social theory*, Harvard University Press, 1994.

Dahlerup, Drude and Lenita Freidenvall, "Quotas as a fast track to equal representation for women: Why Scandinavia is no longer the model," *International Feminist Journal of Politics*, 2005, *7* (1), 26–48.

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

Duffy, John and Jonathan Lafkyz, "Living a Lie: Laboratory Evidence on Public Preference Falsification," *Working Paper*, 2018.

Egorov, Georgy and Bård Harstad, "Private politics and public regulation," *The Review of Economic Studies*, 2017, *84* (4), 1652–1682.

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

_ , “Norms of revenge,” *Ethics*, 1990, *100* (4), 862–885.

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.

Fernández, Raquel, “Cultural change as learning: The evolution of female labor force participation over a century,” *American Economic Review*, 2013, *103* (1), 472–500.

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

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

Goeree, Jacob K. and Leeat Yariv, “Conformity in the Lab,” *Journal of the Economic Science Association*, 2015, *1* (1), 15–28.

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.

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.

Hopfensitz, Astrid and Ernesto Reuben, “The importance of emotions for the effectiveness of social punishment,” *The Economic Journal*, 2009, *119* (540), 1534–1559.

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.

Jones, Stephen R., *The economics of conformism*, Blackwell, 1984.

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.

Kuran, Timur, "The inevitability of future revolutionary surprises," *American Journal of Sociology*, 1995, *100* (6), 1528–1551.

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.

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

Oliver, Pamela E and Gerald Marwell, "Whatever happened to critical mass theory? A retrospective and assessment," *Sociological Theory*, 2001, *19* (3), 292–311.

Oliver, Pamela, Gerald Marwell, and Ruy Teixeira, "A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action," *American Journal of Sociology*, 1985, *91* (3), 522–556.

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

_ **, James Walker, and Roy Gardner**, "Covenants with and without a sword: Self-governance is possible," *American Political Science Review*, 1992, *86* (2), 404–417.

Posner, Eric A., *Law and social norms*, Harvard University Press, 2000.

Schelling, Thomas C, "Micromotives and Macrobehavior WW Norton & Company," *New York, NY*, 1978.

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.

_ , "Social Norms," in S. Durlauf and L. Blume., eds., *The New Palgrave Dictionary of Economics, Second Edition.*, London, Macmillan, 2008.

_ , "The Evolution of Social Norms," *Annual Review of Economics*, 2015, *7* (1), 359–387.

A Proof of Proposition 1

The game has two phases. The first phase lasts from $t = 0$ to $t = t_1$, when type G players initiate the war of attrition. For the analysis of the war of attrition, we use similar arguments as in the proof of Proposition 2 in Egorov and Harstad (2017). In particular, we invoke their result that players' strategies are linear in μ_θ given the payoff structure (see their Appendix B). If type G plays linear strategy $\tau_G(\mu_G) = \frac{\mu_G}{\phi_G \lambda_G}$, the probability that type G has given up by time τ is $\Pr(\mu_G < \phi_G \lambda_G \tau) = 1 - \mathrm{e}^{-\phi_G \tau}$ where ϕ_G is the rate at which type G concedes. For type B, the cost of a conflict of duration τ equals

$$z_B(\tau, t_1) = \int_0^\tau \frac{(1 - \mathrm{e}^{-\gamma t_1})^2 (n-1) p}{\mu_B / x} \mathrm{e}^{-rx} dx.$$

The expected payoff of type B if conceding after duration τ is therefore

$$W_B(\tau, t_1) = \int_0^\tau -z_B(x, t_1) \phi_G \mathrm{e}^{-\phi_G x} dx + \mathrm{e}^{-\phi_G \tau} \left(-z_B(\tau, t_1) - \frac{v_B}{r} \mathrm{e}^{-r\tau} \right)$$

where the first term captures the case when type B is winning the war of attrition and the second term the case when type B is conceding first. Differentiating with respect to τ, we get,

$$\frac{\mathrm{e}^{-\tau(r+\phi_G)} \left(v_B \mu_B (r + \phi_G) - (1 - \mathrm{e}^{-\gamma t_1})^2 (n-1) p r \tau \right)}{r \mu_B}.$$

The first-order condition is therefore satisfied at

$$\tau_B(\mu_B, t_1) = \frac{v_B \mu_B (r + \phi_G)}{(1 - \mathrm{e}^{-\gamma t_1})^2 (n-1) p r}. \tag{5}$$

This is a linear strategy. Further, it can be checked that the second derivative of $W_B(\tau, t_1)$ at $\tau_B(\mu_B, t_1)$ is negative, so this is a global maximum. Similarly, for type G, the cost of a conflict of duration τ equals

$$z_G(\tau, t_1) = \int_0^\tau \frac{\mathrm{e}^{-2\gamma t_1} (n-1) p}{\mu_G / (x n^{TP})} \mathrm{e}^{-rx} dx$$

and the expected payoff of type G if conceding after duration τ is

$$W_G(\tau, t_1) = \int_0^\tau -z_G(x, t_1) \phi_B \mathrm{e}^{-\phi_B x} dx + \mathrm{e}^{-\phi_B \tau} \left(-z_G(\tau, t_1) - \frac{v_G}{r} \mathrm{e}^{-r\tau} \right).$$

Differentiating with respect to τ, we get,

$$\frac{\mathrm{e}^{-\tau(r+\phi_B)}\left(2v_G\mu_G(r+\phi_B)-\mathrm{e}^{-2\gamma t_1}(n-1)pn^{TP}r\tau\right)}{r\mu_G}.$$

The first-order condition is satisfied at

$$\tau_G(\mu_G,t_1)=\frac{v_G\mu_G(r+\phi_B)}{\mathrm{e}^{-2\gamma t_1}(n-1)pn^{TP}r}=\frac{2v_G\mu_G(r+\phi_B)}{\mathrm{e}^{-2\gamma t_1}(n-1)((n-1)p-v_G)r}. \tag{6}$$

In line with the initial assumption, this is a linear strategy. Further, the second derivative of $W_G(\tau,t_1)$ at $\tau_G(\mu_G,t_1)$ is negative, so this is a global maximum. Now, if all players follow these best-response strategies, the concession rates of each type (as expected by the other type) are $\phi_B=\frac{1}{\tau_B(\lambda_B,t_1)}$ and $\phi_G=\frac{1}{\tau_G(\lambda_G,t_1)}$, or after plugging in and rearranging, they are determined by the two equations

$$\phi_B(\phi_G+r)=\frac{(1-\mathrm{e}^{-\gamma t_1})^2(n-1)pr}{v_B\lambda_B} \tag{7}$$

$$\phi_G(\phi_B+r)=\frac{\mathrm{e}^{-2\gamma t_1}n^{TP}(n-1)pr}{v_G\lambda_G}=\frac{\mathrm{e}^{-2\gamma t_1}(n-1)((n-1)p-v_G)r}{2v_G\lambda_G}. \tag{8}$$

Equation (7) defines a hyperbola with asymptotes $\phi_B=0$ and $\phi_G=-r$. Equation (8) defines a hyperbola with asymptotes $\phi_G=0$ and $\phi_B=-r$. These hyperbolas have one intersection with positive (ϕ_B,ϕ_G) and one with negative (ϕ_B,ϕ_G), which proves uniqueness of the equilibrium once the war of attrition has been initiated (the positive solution).

Further, the probability of change (the probability that type B concedes first) is given by

$$\int_0^\infty \phi_G\mathrm{e}^{-\phi_G t}\int_0^t \phi_B\mathrm{e}^{-\phi_B\tau}d\tau dt=\int_0^\infty \phi_G\mathrm{e}^{-\phi_G t}(1-\mathrm{e}^{-\phi_B t})dt=\frac{\phi_B}{\phi_B+\phi_G}. \tag{9}$$

The expected duration of the war of attrition is

$$\int_0^\infty td(1-\mathrm{e}^{-(\phi_B+\phi_G)t})=\int_0^\infty t(\phi_B+\phi_G)\mathrm{e}^{-(\phi_B+\phi_G)t}dt=\frac{1}{\phi_B+\phi_G}. \tag{10}$$

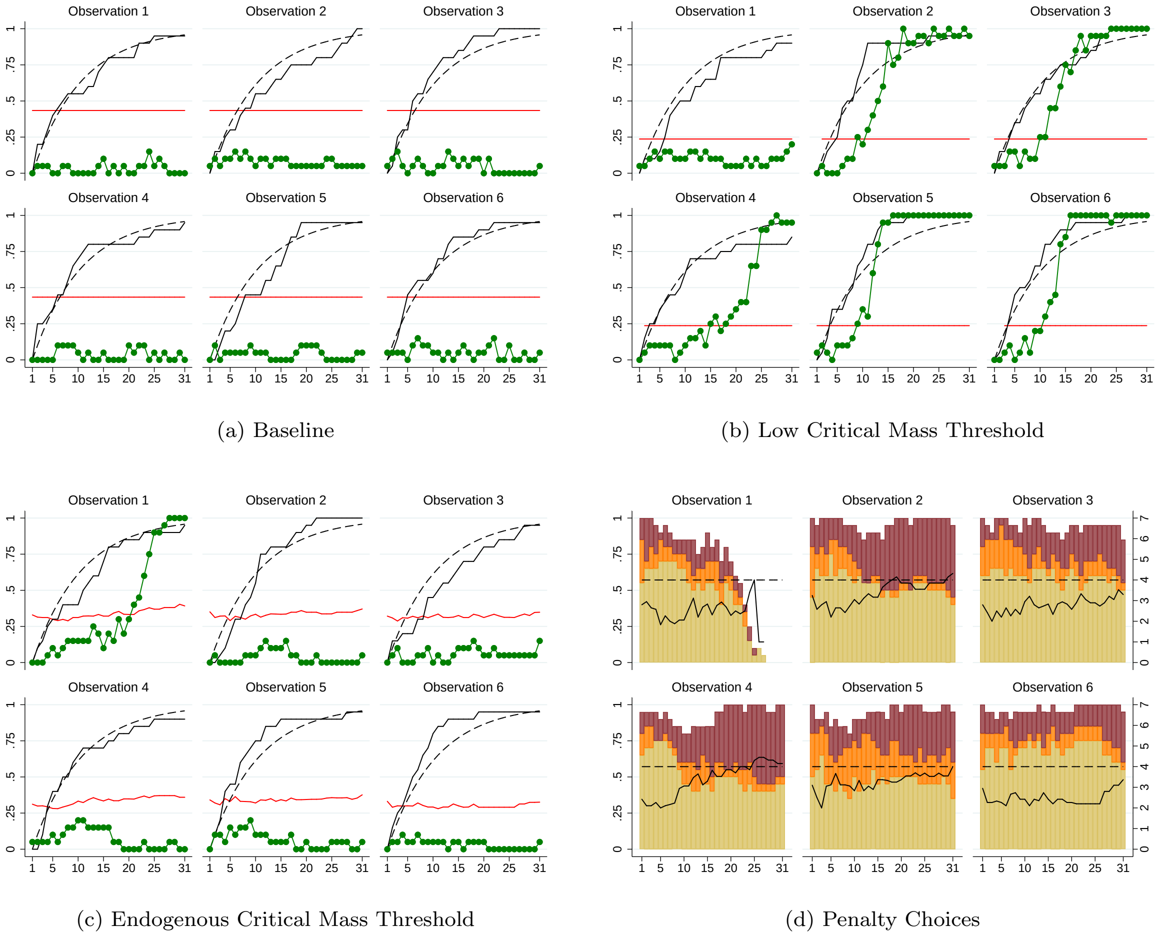
When choosing t_1, the point in time at which to start the conflict, type G players anticipate the equilibrium of the war of attrition. Specifically, type G players choose t_1 to maximize

$$\int_0^{t_1}-\mathrm{e}^{-rx}v_G dx+\mathrm{e}^{-rt_1}W_G(\tau_G(\lambda_G),t_1). \tag{11}$$

The optimal time t_1 is reached when the marginal benefit of delaying t_1 and increasing type G's winning chances in the war of attrition (derivative of second term) falls short of the marginal cost of choosing the less preferred action for another time unit (derivative of first term). This problem has a generically unique solution.

Note on (7) and (8):

- An increase in v_B or λ_B shifts the first hyperbola down and left and does not affect the second



(a) Baseline

(b) Low Critical Mass Threshold

(c) Endogenous Critical Mass Threshold

(d) Penalty Choices

Figure 9: ...

Notes: Circled markers represent the fraction of subjects choosing *green*. The solid increasing line shows the fraction of subjects preferring *Green*; the dashed line the corresponding theoretical expectation. The horizontal line is the Critical Mass Threshold.

one. The intersection point therefore moves along the second hyperbola up and to the left. Thus, ϕ_B decreases and ϕ_G increases. For analogous reasons, an increase in v_G or λ_G leads to a decrease in ϕ_G and an increase in ϕ_B.

- An increase in γ or t_1 shifts the first hyperbola up and right and the second hyperbola down and left. Thus, ϕ_B increases and ϕ_G decreases.
- Holding the maximal disunity penalty $(n-1)p$ constant (i.e., decreasing p as we increase n and vice versa), an increase in n shifts the second hyperbola up and right and does not affect the first one. This implies an increase in ϕ_G and a decrease in ϕ_B.