

Living a Lie: Theory and Evidence on Public Preference Falsification*

John Duffy[†] Jonathan Lafky[‡]

August 3, 2018

Abstract

We propose a model of how public behavior changes in response to the evolution of privately held preferences. Our aim is to rationalize the tendency for individuals who hold minority viewpoints to falsely report their preferences by taking actions favored by the majority. We do this using a game involving a tension between honest expression of one's true preferences and a desire to conform to the behavior of others. In an experimental test of our model, we find confirmatory support for the model's main predictions, that even after a majority of the population shares what was previously an unpopular minority opinion, a lack of mutual awareness among members of the new majority can allow continued public support for the old status quo, and that the onset and speed of transitions to new, majority-held opinions depend on the relative difference in rewards from conformity versus truthful expression.

Keywords: Conflict, Conformity, Social Change, Hypocrisy, Insincerity, Groupthink, Pluralistic Ignorance, Preference Falsification, Experimental Economics.

JEL Numbers: C92, D74, D82, D83.

*Funding for this project was provided by the UC Irvine School of Social Sciences. We thank Tyler Boston for expert research assistance.

[†]Department of Economics, University of California, Irvine, CA 92697. Email: duffy@uci.edu Phone: (949) 824-8341

[‡]Carleton College Department of Economics, Northfield, MN 55057. Email: jlafky@carleton.edu Phone: (507) 222-4103

1 Introduction

Individuals often allow concerns for conformity to take precedence over their true, private preferences. Some examples of this behavior are mundane, such as dressing in a less-preferred style because it is common among friends, or ordering a beer rather than a preferred piña colada so as to fit in with colleagues on a Friday night. Other examples are more impactful, such as a person expressing public opinions about issues like same-sex marriage or drug legalization that are at odds with their own privately held beliefs. At the extreme, such behavior could lead many to avoid participating in an anti-government revolution that they otherwise believed in, out of a concern that the revolution would fail because of insufficient popular support, e.g., citizens of communist regimes in the mid-20th century. In each of these examples, there may in fact be many people who simultaneously chose their less-preferred action in order to conform with those around them. Everyone at the bar might secretly prefer a piña colada, but continue to order the socially acceptable beer, while every person in a society might prefer to overthrow the dictator, but take no action to avoid being ostracized (or worse).

While such behavior has been described in many ways, e.g., insincerity or hypocrisy, we adopt Kuran’s (1995) terminology of “public preference falsification” to refer to situations in which large numbers of individuals publicly express a position that is opposite to their own, true, privately held position.¹ Kuran argues that there are two features of preference falsification that differentiate it from other forms of lying. First, preference falsification brings disutility to the falsifier. Thus, so-called “white lies” – falsities delivered with good intentions – may be excluded. Second, preference falsification occurs out of a desire to conform to the real or imagined preferences of the plurality of others. Thus, preference falsification is distinct from *self-deception* as it requires a demonstration of *public* behavior that is distinct from one’s own true preference out of a desire to conform to the preferences of the plurality. In other words, preference falsification is a *social phenomenon*. Importantly, the preferences of the majority may be imperfectly known, and thus perceptions of it may be incorrect, a situation that social psychologists refer to as “pluralistic ignorance” (Allport 1924).² The consequences of pressures to conform in the face of pluralistic ignorance can be significant, and include, for example, support for policies of racial oppression and segregation (O’Gorman 1975), acceptance of binge

¹The notion that there could be widespread preference falsification predates Kuran, of course. For instance, it is lampooned in Hans Christian Anderson’s fairy tale, “The Emperor’s New Clothes,” Anderson (1838). In modern parlance, reference is often made to the unacknowledged “elephant in the room” referring to groupthink that nobody is willing to challenge.

²Pluralistic ignorance arises when individuals mistakenly believe that they hold minority beliefs. They conform to social norms they do not privately support but believe that others’ sincerely believe in those social norms. Game theoretically, pluralistic ignorance implies an absence of common knowledge. (Chwe 2001).

drinking on college campuses (Prentice and Miller 1993) and tolerance of smoking by others (Sherman et al. 1983), despite majorities opposed to such activities.

Preference falsification can be seen as both a cause and effect of coordination failure, where the outcome that individuals have coordinated upon is not Pareto efficient. A desire to conform to the majority action can prevent an individual from revealing his or her own true preference, thereby helping to sustain an old equilibrium outcome even after a majority have switched to preferring an alternate outcome. Coordination failures in turn lead to falsification when an inefficient equilibrium persists even after preferences have evolved in favor of the new outcome. In other words, individuals hiding their true preferences effectively sustain existing inefficient equilibria, and the sustained inefficient equilibrium causes individuals to avoid revealing their true preferences.

This paper is not the first to study public preference falsification or pluralistic ignorance - see the literature review for some relevant key references. Rather, the contribution of this paper is that we use the tools of game theory and experimental economics to model and to empirically evaluate the public preference falsification phenomenon. Kuran (1995) discusses how one might measure preference falsification and what it would take to challenge the preference falsification theory. He suggests that, like Darwin, we should look to historical, natural experiments for supportive evidence, but he also notes that “natural experiments are seldom precise enough. Their power is often diminished by factors that one would have wanted to hold fixed.” (Kuran 1995, p. 343).

Recognizing this control problem, we propose to model the preference falsification phenomenon as an n -player, non-cooperative game and to explore how players play that game in a laboratory experiment. Specifically, we study an environment where players are induced to hold private preferences about some binary issue and must express a public preference about that same issue. Each player’s utility depends on the extent to which their public preference conforms with the preferences of the other $n - 1$ players, but also on the extent to which their public preference is consistent with their own private preference or “type.” Specifically, a player’s utility from expressing a public preference is increasing with the number of others expressing that same preference, but expressing a public preference that is at odds with one’s own private preference (type) brings disutility to the player. Further, in our framework, players’ preferences evolve over time. Initially, all players’ private preferences are the same. Over time, some players’ preferences change from the original preference to the opposite preference. For example, initially all players might hold the preference that cigarette smoking is acceptable behavior in public spaces. Over time, due to (say) evidence that smoking causes lung cancer, players switch their preference to opposing smoking in public spaces. This evolution

of preferences over time is a key feature of our approach that enables us to identify whether, and to what extent, public preferences may depart from private preferences.

The main problem with studying preference falsification in the field is that it may not be possible to know individuals' true private preferences at a given moment in time. Our approach of inducing private preferences allows for careful control over those preferences, albeit with the usual caveat that induced preferences can only approximate those in natural settings of social conformity that we seek to model. By inducing players to hold certain preferences and by being neutral about the decision-making context, we believe that we can more clearly test whether individuals are being truthful in the public statement of their preferences. Indeed, our inducement of preferences, the neutral framing of the choice task, and the monetary incentives that we provide are the main advantages that our approach offers over other approaches to studying preference falsification. In addition, unlike social psychology experiments studying conformism, beginning with the seminal work of Asch (1956), we do not require the use of confederates and thus avoid the need to engage in any type of deception. Instead, we implement a non-cooperative game similar to an n -player Battle of the Sexes games, but where the induced preferences over outcomes gradually change over time.

In developing a simple n -player game and evaluating how players play that game in the laboratory, we ask several related questions. First, does preference falsification occur, and if so, can it be sustained after a majority of group members have abandoned the old preference? In other words, can the majority in a group act against their own self interest due to historical behavior? We find in our experiment that this phenomenon is uncommon, but can happen when the payoffs from conformity are sufficiently high. Second, we ask how sensitive the frequency and duration of preference falsification is to the relative size of the private and public incentives for behavior. Our theory predicts that, on the one hand, a larger incentive to conform results in greater preference falsification and delays the onset of the transition to the new social norm. On the other hand, when incentives to conform are greater, our theory also predicts that the transition, if and when it occurs, will be more abrupt than when incentives to conform are lower. Finally, we ask whether greater uncertainty about the majority position - greater pluralistic ignorance - affects the timing of the transition to the new social norm. We find substantial support for our predictions in the experimental data.

2 Related literature

The notion of public preference falsification and its social consequences was first elaborated upon by Kuran (1995), though social psychologists have long studied the question of social norm compliance in the face of pluralistic ignorance, see, e.g., Moscovici (1985), and Turner

(1991). Experimental evidence for conformity in group processes was first presented by Asch (1956) using a line judgment task, where confederates of the experimenter exerted pressure on subjects to conform to their mistaken judgments of the length of a line. A key difference between this work and the present paper is that we are using both game theory and experimental economics methods to study the problem, and thus we avoid any type of deception.

Indeed, the problem we study can be viewed as a coordination game with heterogeneous preferences, as in Battle of the Sexes, of which there exist several experimental studies, including Cooper et al. (1989, 1993), Charness et al. (2007), and Crawford et al. (2008) among others; Ochs (1995) and Devetag and Ortmann (2007) provide surveys of this experimental coordination game literature. A key difference between the environment we study and prior experimental studies of coordination games is that we consider $n > 2$ -player versions of such games where the payoff incentives of the game change over time with the change in players' preferences or types.

In economics, social influence has been mainly studied in the context of the information herding models of Banerjee (1992) and Bikhchandani et al. (1992), where individuals may rationally ignore their own private information in favor of following the choices made by predecessors in their objective of forecasting the true but unknown state of the world. Experimental support for the rational herding phenomenon was first provided by Anderson and Holt (1997). Hung and Plott (2001) replicate Anderson and Holt's findings and study a modified information cascade game where players are rewarded both for guessing the correct state of the world and well as for conforming to the decisions made by the majority of others, which bears some resemblance to the payoff structure of the game we study. Hung and Plott report that the addition of incentives for conformity reduce efficiency relative to the baseline individual payoff case. Goeree and Yariv (2015) show that some subjects (34%) prefer to observe uninformative histories of play from other subjects, rather than their own informative private signal. The subjects who choose to view such histories tend to take the same action as their predecessors, and Goeree and Yariv argue that this is evidence of an innate preference for conformity, independent of explicit incentives for either conformity or accuracy.

Despite some similarities, there are important differences between our environment and the information herding model. First, while our game is repeated, the decisions of all players are made simultaneously (and not sequentially) so that players cannot condition their choice for the current period upon the choices made by others in that period. Second, and most significantly, in our setting players are not seeking to identify the true and common state of the world; rather they are asked to express a public preference, the payoff consequences of which depend on their own private preference and the publicly expressed preferences of

others. By contrast with the information herding game, the private preferences of individuals in our model change over time. The herding behavior in our environment does not result from observing others' best guesses about an unchanging state of the world. Rather, herding in our environment results from a desire for conformity, and on the resultant impact of that desire on individuals' beliefs about the social norm, which can conflict with their own true preferences.

There is also a related experimental literature suggesting that pro-social behavior, e.g., giving to charities, may be driven by a reluctance to depart from the perceived beliefs of others regarding appropriate behavior, as opposed to a true individual preference for giving, e.g., Dana et al. 2007, DellaVigna et al. 2012. By contrast, we study a neutrally framed coordination game under evolving preferences, where there is no meaningful labeling of the different actions that players can take so that prior notions of what is or is not "pro-social" are not operational.

Michaeli and Spiro (2015, 2017) have developed theoretical models of conformity to social norms in a *heterogeneous* agent framework where norm compliance is a continuous choice variable. In their framework, each agent's total loss is the sum of separable private discomfort and social pressure components. They use their model to understand the different patterns of norm conformity across societies and to show how norms can be biased relative to average preferences, findings that depend on the parameterization of their model. While we also consider a heterogeneous agent framework in studying norm compliance, we consider the simpler case of binary adherence to a social norm (or not) and our main contribution is that we implement and evaluate our model in the experimental laboratory.

In concurrent research, Andreoni et al. (2017) also study social change in the laboratory with evolving private preferences. While they focus on probabilistically evolving preferences, we compare and contrast environments in which the change in preferences is deterministic or probabilistic. Also differently from our study, in the Andreoni et al. design players are matched, payoffs are assessed and information is revealed pairwise, whereas in our setting, payoffs depend on group decisions (i.e. we employ n -player matching), as we have in mind that actions are publicly revealed. As in our study, Andreoni et al. find, in their baseline treatment, that subjects can get caught in what they call a "conformity trap", which amounts to all subjects choosing an action that is different from their privately preferred action. We observe fewer instances of such conformity traps in our design, which involves a different payoff function that does not vary across treatments as in Andreoni et al.'s design. We view our study as complementary to theirs, especially concerning knowledge about the population-wide distribution of types or actions.

3 Theory

In this section we outline a theoretical model that we will later test in the laboratory. In our model (as well as in our experiment), at any moment in time, individuals can be one of two possible types. We consider two environments for the evolution of types over time. In the first environment there is a known, deterministic process by which each individual's type changes over time. The second environment is identical, except that the process by which an individual's type changes over time is stochastic. We make the simplifying assumption that changes in type occur only once and remain permanent, meaning that once an individual has switched their type, they never revert back to their original type. This assumption makes the predictions of our model as clear as possible.³

3.1 Deterministic Progression of Types

The game consists of n players, each of whom makes an action choice in each of the $t = 1, 2, \dots, T \geq n$ periods of the game. At the start of each period, each player i has a *private* type, $\theta_i \in \{X, Y\}$, that is known only to themselves. With knowledge of their own type, each player i simultaneously chooses a *publicly* observable action, $a_i \in \{X, Y\}$. Player i 's stage game (period) payoffs are given by:

$$U_i(a_i, \theta_i, k_{a_i}) = \begin{cases} H \cdot \frac{k_{a_i}}{n} & : a_i = \theta_i \\ L \cdot \frac{k_{a_i}}{n} & : a_i \neq \theta_i \end{cases}$$

where $H > L$ and k_{a_i} is the total number of players who choose action a_i in that period (stage) of the game. We assume that $H \frac{1}{n} < L \frac{n}{n}$, or equivalently, $H < Ln$, thereby excluding the trivial case in which every player simply takes their preferred action, regardless of the actions of others.

In period $t = 1$ all subjects have the same private type, $\theta_i = X$. Each period thereafter, exactly *one* player's private type switches from $\theta_i = X$ to $\theta_i = Y$. These features are common knowledge among all n players. In other words, in period $t = 1$ the count of X -type players is $c_1^X = n$ and the count of Y -type players is $c_1^Y = 0$. In period $t = 2$ the counts are $c_2^X = n - 1$ and $c_2^Y = 1$, and more generally, in period $t = j$ there are $c_j^X = \max\{n + 1 - j, 0\}$ X -type players and $c_j^Y = \min\{j - 1, n\}$ Y -type players. We chose this deterministic process in the interest of simplicity and clarity; in this setting, agents can perfectly forecast the majority type in any period of the game, so that "pluralistic ignorance" is minimized. In the next

³It is also not unreasonable. For instance, politicians and others who repeatedly flip their positions back and forth are referred to, with some opprobrium, as "wafflers" (an analogy to an easy-to-flip breakfast food).

section we consider a stochastic transition process, which has the effect of creating pluralistic ignorance.

The transition of types from all X to all Y can be given several interpretations. Our preferred interpretation is to view the transition of types as a gradual change in individual tastes or preferences on some social issue, e.g., the gradual evolution from acceptance to rejection of smoking in public. Alternatively, the transition of types might reflect changing knowledge or understanding, e.g., producers in an industry gradually discovering the superiority of one production technique over another.

Regardless of the distribution of types in any period of the game, all players choosing X or all players choosing Y will always comprise Nash equilibria. There are also equilibria that exhibit honest disagreement, in which some players choose X while others choose Y . We focus attention on the progression from an initial all- X equilibrium, to an disagreement equilibrium in which all players take an action matching their type, to an eventual all- Y equilibrium, as this is the most efficient equilibrium path of the game.

Consider a Y -type who is currently choosing action $a_i = Y$. The Y -type prefers to play Y if it yields a higher payoff than $a_i = X$, or equivalently, when the payoff from choosing Y , or $H \cdot \frac{k_Y}{n}$, is greater than the payoff from switching to playing X instead, or $L \cdot \frac{k_X+1}{n}$. Since $k_X = n - k_Y$, this condition is equivalent to $k_Y > \frac{(n+1)L}{H+L}$. In other words, $k_Y > \frac{(n+1)L}{H+L}$ is a necessary condition for an equilibrium with disagreement. Prior to this threshold being met, Y -types constitute too small a minority to support choosing $a_i = Y$, even if all Y -types did so simultaneously. Similarly, an X -type player should choose as their public action $a_i = Y$ if it results in $L \cdot \frac{k_Y}{n} > H \cdot \frac{k_X+1}{n}$, or equivalently $k_Y > \frac{(n+1)H}{H+L}$. The condition $k_Y > \frac{(n+1)H}{H+L}$ describes when an equilibrium with disagreement is no longer sustainable, as X -types will prefer to switch to reporting $a_i = Y$, resulting in the unanimous choice of Y among all players.

Let $\tau_1(H, L, n)$ denote the smallest number of Y -types such that $k_Y > \frac{(n+1)L}{H+L}$ if all Y types play $a_i = Y$, and let $\tau_2(H, L, n)$ denote the smallest number of Y -types such that $k_Y > \frac{(n+1)H}{H+L}$ if all X types play X . Note that under the deterministic progression of types, the threshold τ_1 occurs in period $\tau_1 + 1$ while τ_2 occurs in period $\tau_2 + 1$, and that $\tau_1 < \tau_2$. It follows that there exists a subgame perfect equilibrium in which $a_i = X$ for all i until τ_1 is reached followed by $a_i = \theta_i$ (a period of “disagreement” or “honesty”) during the transition phase, when the number of Y -type players ranges from τ_1 to τ_2 , and finally $a_i = Y$ for all i once τ_2 is reached. These predictions are illustrated in Figure 1.

As Figure 1 makes clear, there are two sets of players who are predicted to “live lies” in this equilibrium - those whose private type switches from type X to type Y in the first $\tau_1 + 1$ periods but who publicly announce X during those same periods and those who privately

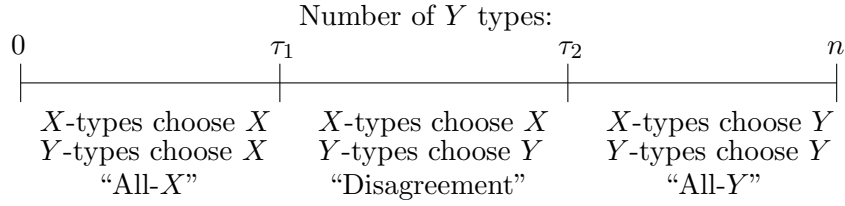


Figure 1: Predicted Public Choices, Given the Number of Y types

remain type X in the periods starting from period $\tau_2 + 1$ onward and who publicly announce Y in those periods. The first types are those who might be termed “ahead of their time” who hold initially unpopular preferences that later become the norm, while the second type might be regarded as “old-fashioned”, retaining now-unpopular preferences that were once commonplace.

To better illustrate the theoretical predictions, we present two numerical examples that are also used in our experiment. Let $n = 12$, $H = 9$ and $L = 3$. Then, Y -type players will switch from $a_i = X$ to $a_i = Y$ if making the switch results in $k_Y > \frac{(n+1)L}{L+H} = \frac{(12+1) \cdot 3}{9+3} = 3.25$, meaning that $\tau_1 = 4$, which occurs in period $\tau_1 = 5$. At this point, if all Y -type players switch to action Y , we have $k_Y = 4$, and $k_X = 8$, as all remaining X -type players are still willing to choose $a_i = X$. The remaining X -type players keep choosing X (honestly) until $k_Y > \frac{(n+1)H}{L+H} = \frac{(12+1) \cdot 9}{9+3} = 9.75$, meaning that $\tau_2 = 10$, which occurs in period $\tau_2 + 1 = 11$. From period 11 onward, any and all remaining X -type players will switch to taking public action Y .

Alternatively, suppose that $n = 12$, $H = 9$ and $L = 6.43$. In this case the ratio between the two payoffs, H and L , is not as great, so the pressure to conform to the majority opinion will be greater than in the previous example. With this new parameterization, a Y -type player will choose action Y if it results in $k_Y > \frac{(n+1)L}{L+H} = \frac{(12+1) \cdot 6.43}{9+6.43} = 5.42$, meaning that $\tau_1 = 6$, which occurs in period $\tau_1 + 1 = 7$. Both Y and X -types will report honestly until the second threshold is crossed, when $k_Y > \frac{(n+1)H}{L+H} = \frac{(12+1) \cdot 9}{9+6.43} = 7.58$, meaning that $\tau_2 = 8$, which occurs in period $\tau_2 + 1 = 9$. In other words, beginning in period 9, the remaining X -types switch over to taking public action Y .

This numerical example illustrates two features of the subgame perfect equilibrium that we will test in our experiment. First, when the ratio H/L is large, so that the incentives to conform are low, the *onset* of the transition, i.e., the first critical proportion of Y -types, τ_1 , occurs *earlier* in time than when the ratio H/L is smaller so that the incentives to conform are greater. Second, when the ratio H/L is large, the *transition* phase, or the period of “disagreement,” as defined by the difference $\tau_2 - \tau_1$, is *longer* than when the ratio H/L is

smaller; in the latter case the transition is relatively more abrupt.

3.2 Probabilistic Progression of Types

In many scenarios involving public preference falsification, there is uncertainty as to the proportion of the population that prefers each action, i.e., there is true pluralistic ignorance. We make the simplest change to our deterministic model in order to capture this uncertainty. Specifically, we retain the transition of types over time as in the deterministic model, but instead of a steady and perfectly known increase in the number of Y -types each period, the number of Y -types is only stochastically increasing in each period. As in the deterministic case, we allow for at most one player to switch type in each period following the first period and we maintain the assumption that players only switch from type X to type Y , never reverting back from type Y to type X .

More precisely, we assume that at the start of each new period $t > 1$ with a commonly known constant probability, p , exactly one player, among the set of X -type players as of period $t - 1$, switches from being an X -type player to being a Y -type in period t and remains a Y -type in all subsequent periods, $t + 1, \dots, T$. In this case, the expected number of Y -type players in period $t = 1$ is 0, the expected number in period $t = 2$ is p , the expected number in period $t = 3$ is $0(1 - p)^2 + 2p(1 - p) + 2p^2 = 2p$, and so on. Recognizing that the number of Y -type (or X -type) players can never exceed n , we have that the expected count of Y -types in period t , $E[c_t^Y] = \min\{p(t - 1), n\}$ and likewise, the expected count of X -types is given by $E[c_t^X] = \max\{n - p(t - 1), 0\}$. This change has the effect of increasing uncertainty (or the degree of pluralistic ignorance) about the majority type relative to the deterministic transition case.

In the deterministic case, the number of Y -type players in period t is strictly greater than the corresponding *expected* number of Y -type players in period t in the probabilistic case. Thus, assuming that players can correctly form expectations about the number of each player type, the transition from a public choice of X by all n players to a public choice of Y by all n players should come later in time in the probabilistic case as compared with the deterministic case; how much later in time will depend on the precise choice of p .

4 Experimental Design

We implement a 2×2 experimental design, where one treatment variable is the incentive to conform, low conformity (LC) or high conformity (HC), and the other treatment variable is the nature of the change in player types, deterministic (D) or probabilistic (P). Thus our four

treatments are: deterministic low conformity (DLC), deterministic high conformity (DHC), probabilistic low conformity (PLC) and probabilistic high conformity (PHC). Across all four treatments we hold constant the number of players in each group, $n = 12$, the total number of periods in each game, $T = 20$ and the number of games played, 2.

In the low conformity treatments, the maximum payoff from taking an action that matches a subject's type is $H = \$9.00$, while the maximum payoff for taking an action that differs from a player's type is $L = \$3.00$. In the high conformity treatments, by contrast, the matching payoff remains $H = \$9.00$, while the mismatching payoff, $L = \$6.43$. These parameterizations were chosen in order to test the comparative statics predictions of the theory, while providing distinct predictions between treatments for the onset, duration and completion of the transition from the all- X equilibrium to the all- Y equilibrium.⁴ As discussed earlier in the numerical examples of Section 3.1, in the LC treatments, $\tau_1 = 4$ and $\tau_2 = 10$. By contrast, in the HC treatments, $\tau_1 = 6$, and $\tau_2 = 8$, so in the HC treatments, the transition phase starts later but has a shorter duration relative to the LC treatments.

In both the deterministic and probabilistic treatments, all $n = 12$ players begin period 1 as X -types. In the deterministic treatment, exactly one X -type player, from all remaining X -type players switches to being a Y -type in each period $t = 2, 3, \dots$ until all 12 of the X -types have switched to Y -types in the 13th period of the $T = 20$ period game. This transition pattern is carefully explained to subjects in the written instructions and can therefore be viewed as public knowledge. In the probabilistic treatment, we set $p = 0.75$ so that there is a 75% chance of exactly one remaining X -type permanently switching to being a Y -type in each period $t > 1$ up until the point that all $n = 12$ subjects are Y -types, after which no further switching of types takes place. The choice of $p = 0.75$ is made known to all subjects in the written instructions of the probabilistic treatment sessions, and can thus be regarded as public knowledge. With $p = 0.75$, in expectation, all players should be Y -types by period 16 of the 20 period game. Our choice of $p = 0.75$ is motivated by the desire to make it very likely that a full transition from all X -types to all Y -types is likely to have occurred by the final, 20th period of the game. Indeed, we conducted a simulation analysis of our game with $n = 12$ subjects and $p = 0.75$, and found that over 100,000 trials, the median number of X (Y) types by the final, 20th period, is 0 (20) with 75 percent confidence. Thus we had confidence that the type progression would be complete by the 20th period.

In the probabilistic treatments, each group $i = 1, 2, \dots, 6$ of treatment PLC experiences an independent, randomly determined sequence of type progressions using $p = 0.75$ in each

⁴The H and L payoffs were chosen such that Y -type subjects received the equal payoffs from playing either X or Y when there were an integer number (5 in HC, 3 in LC) of Y -types choosing Y .

of parts 1 and 2. We then use those *same* random sequences of type progressions for parts 1 and 2 for one matched group $i = 1, 2, \dots, 6$ of the PHC treatment. Thus, each group in treatment PLC has one matched group in treatment PHC that experiences the exact *same* random sequence of type progressions in parts 1 and 2. Pairing groups across treatments in this manner allows us to minimize the effect of randomness on differences in observed behavior between the two treatments, while still allowing some variation in the probabilistic realizations that our subjects face, namely 6 different random sequences of type progressions in each part of the experiment. All of the experimental parameters are summarized in Table 1.

Table 1: Experimental Parameters

	Treatment			
	DLC	DHC	PLC	PHC
Groups	6	6	6	6
Subjects	72	72	72	72
Periods	40	40	40	40
Type progression	Deterministic	Deterministic	Probabilistic	Probabilistic
Conformity payoff	\$9.00	\$9.00	\$9.00	\$9.00
Non-conformity payoff	\$3.00	\$6.43	\$3.00	\$6.43
Y -threshold, τ_1^a	4 [5]	6 [7]	4 [7]	6 [9]
X -threshold, τ_2^a	10 [11]	8 [9]	10 [15]	8 [13]

^aAs defined in Section 3, the switch thresholds τ_1 and τ_2 are the minimum number of Y -types in the population of size 12 that are needed for subjects of a given type (Y -type for τ_1 , X -type for τ_2) to switch from playing X to Y . The time period in which this switch threshold occurs (deterministic) or is expected (probabilistic) is shown in brackets [].

Table 1 also reports the various switching thresholds, given our parameterization of the game and the period in which each switch is predicted to take place (or can be expected to take place). For instance, in the deterministic, low conformity (DLC) treatment, when there are at least 4 Y -types in the population, (which occurs in period 5), there exists a Nash equilibrium in which all X -types play X and all Y -types play Y , i.e., there exists a Nash equilibrium with truthful revelation of types. When there are at least 10 Y -types in the population (which occurs in period 11), this truth-telling Nash equilibrium no longer exists. The thresholds for the other three treatments are also shown. Notice that the period in which these switch thresholds occur (or are expected to occur) are always earlier in time in the deterministic version of a treatment (LC or HC) as compared with the corresponding probabilistic version

of that same treatment.

This study was computerized using the z-Tree (Fischbacher, 2007) software and conducted in the Experimental Social Science Laboratory at the University of California, Irvine. Subjects were undergraduate students with no prior experience playing this game.

Subjects in each session were assigned to groups of size 12, and participated in a total of 40 periods of decision making, broken up into two, 20 period games or “parts”. Subjects were members of the same group of 12 for both parts. They were initially provided with instructions only for the first part of the experiment, though they were informed that there would be two parts to the experiment, and that they would receive additional instructions for the second part after they completed the first part. While the second part was a repeat of the first part, we did not inform subjects of this fact in advance. At the conclusion of part 1, subjects were given continuation instructions for part 2 revealing that the second part would be a repeat of the first part. Subjects were paid their earnings from two randomly selected periods, one drawn from each 20-period part.

At the start of each experimental session, the written instructions were distributed and read aloud in an effort to make the information public knowledge. The instructions avoid reference to preferences, falsification, lying, etc. so as to provide a neutral setting in which to fairly evaluate the theory. Copies of the instructions used in the experiment are provided in the Appendix. Following completion of the instructions, subjects had to correctly answer a number of control questions designed to check their comprehension of the written instructions. Subjects who had incorrect answers were asked to reconsider their choices and the experiment did not commence until all subjects had correctly answered all control questions.

Parts 1 and 2 began identically, with all $n = 12$ subjects starting out as X -types. At the beginning of each period, prior to making any choices, each subject was informed as to their type on their computer screen – they were either an “ X -type” or a “ Y -type.” After viewing this information, they then chose whether to take action X or action Y for the period. After all subjects made their choices, each subject learned their payoffs for the period, as well as the total number of subjects in their group who chose each action (X or Y). The total number of subjects of each *type* was never reported, though it could be easily inferred in the deterministic treatment, and the expected number of each type could be inferred in the probabilistic treatment. The latter choice was made to give a simple representation of pluralistic ignorance.

Following the 20th period, subjects were informed that the first part of the experiment had ended, and that they would now begin a new sequence of 20 periods of play, following the same rules as in the first part and interacting with the same 12 players as in the first part. We

chose to keep subjects in the same matching group of 12 in the second part so as to provide conditions most favorable to learning, which we examine later in the paper. The second part of the experiment began identically to the first, with all subjects reset to being X -types, and the same type progression rules were in effect. However, the *order* in which subjects' types switched from X to Y was randomized independently from the first part, meaning that subjects in part 2 were not aware of *when* their own type would change, as was also the case in part 1.

4.1 Hypotheses

We note first that, as discussed in Section 3, in every treatment and regardless of the distribution of player types at any moment, within each period there always exist two pure strategy Nash equilibria, one in which all players choose action X and another one in which all players choose action Y . Focusing only on individual deviations, it is never individually profitable to switch from a unanimously chosen majority action to being the only person taking the minority action. Therefore, we would expect that a group starting at the natural equilibrium of all players choosing action X would never experience anyone choosing action Y .

If we allow for simultaneous deviations across multiple players (i.e., multi-player coalitions), then switching away from all players choosing action X can become profitable. In the LC treatments, as Table 1 reveals, Y -types are better off by collectively choosing action Y in the low conformity treatment when there are four or more Y -types. In other words, it is not possible for a group of fewer than four Y -types to increase their payoffs by switching away from the all- X equilibrium. Similarly, in the HC treatments, Table 1 reveals that Y -types prefer collectively choosing action Y when there are six or more Y -types. However, when the number of Y -types reaches a second threshold, 10 in the LC and 8 in the HC treatments, X -types now prefer to also switch to collectively choosing action Y . Thus, as noted earlier, in the LC treatments there exists a “disagreement” Nash equilibrium where agents choose the action corresponding to their own type so long as the expected number of Y -types is between 4 and 9. In the HC treatments, there exists a similar disagreement Nash equilibrium where agents choose the action corresponding to their own type so long as the expected number of Y -types is between 6 and 7, inclusive.

With the foregoing analysis in mind, and using payoff efficiency as our equilibrium selection criterion, we posit four hypotheses that we will test with our experimental data:

Hypothesis 1. *Groups switch from the all- X equilibrium to the disagreement equilibrium when a smaller number of subjects are Y -types (and thus earlier in time) in the low-conformity treatments than in the high-conformity treatments.*

Hypothesis 1 is specifically about the empirical relevance of the first critical threshold, τ_1 , marking the earliest possible onset of the transition. The value of τ_1 is lower in the low conformity treatments as compared with the high conformity treatments (see again Table 1). Intuitively, Hypothesis 1 says that subjects with novel preferences “speak up” sooner when the pressure to conform is lower.

Hypothesis 2. *The disagreement phase, in which all players take actions equal to their types, and defined by $\tau_2 - \tau_1$, is longer in the low-conformity treatments than in the high-conformity treatments.*

This hypothesis says that, despite the earlier onset of the transition to the all- Y equilibrium in the low conformity treatment (Hypothesis 1), the intermediate phase of honest disagreement will last longer in the low conformity treatment than in the high conformity treatment. Indeed, as Table 1 reveals, the number of Y -types needed before the transition to the all- Y equilibrium is predicted to be complete, τ_2 , is *higher*, and thus comes later in time, in the low conformity treatment as compared with the comparable high conformity treatment. Intuitively, Hypothesis 2 says that, while higher pressure to conform may delay individuals from acting on new preferences, once those preferences are shared they are rapidly adopted by the group.

Hypothesis 3. *The disagreement phase is shorter and begins earlier in the deterministic treatments than in the probabilistic treatments.*

This hypothesis follows immediately from the fact that $p < 1$ in the probabilistic treatments. As Table 1 reveals, both threshold switching periods, τ_1 , τ_2 , come earlier in time in the deterministic treatment than in the comparable probabilistic treatment. Thus Hypothesis 3 is a simple check on whether the probabilistic treatment delayed the period in which a transition from all X to all Y was initiated.

Because the expected number of Y types in any period differs between the probabilistic and deterministic treatments, we will condition much of our analysis on the number of Y types in the population. Using number of Y -types instead of periods allows for more equivalent comparisons of subject behavior across treatments:

Hypothesis 4. *The number of Y -types at which the disagreement phase begins and ends is the same between deterministic and probabilistic versions of the same conformity treatments, (DLC versus PLC and DHC versus DHC).*

Hypothesis 4 conditions on the number of Y -types in the population and *not* on the period number. The hypothesis follows from the fact that τ_1 and τ_2 are the same between

DLC and PLC and between DHC and PHC (see again Table 1). Intuitively, Hypothesis 4 says that, holding pressure for conformity constant, the tendency to disagree depends only on the popularity of an opinion, regardless of the speed at which the opinion spreads.

5 Results

As noted in Table 1 we report experimental results from six different, 12-player groups for each of our four different treatments, deterministic low conformity (DLC), deterministic high conformity (DHC), probabilistic low conformity (PLC), and probabilistic high conformity (PHC). Our experiment thus involves a total of $6 \times 12 \times 4 = 288$ experimental subjects. Average subject earnings across all four treatments were \$15.01, plus a \$7.00 show-up fee. Each subject's participation in an experimental session lasted about 90 minutes.

In every group, every subject's type had switched from X to Y before the final, 20th period in each part of the experiment. The complete transition of types from all- X to all- Y was assured in the deterministic treatment where it always happened in period 13, but such a complete transition was not ex ante guaranteed in the probabilistic treatments. Table 2 reports the first period number, of each part of the experiment, in which 100% of subjects (all 12) were type Y . Recall that for each probabilistic treatment (PLC or PHC) group, we used the same realization of the probabilistic transition process for *pairs* of groups, one assigned to PLC and one assigned to PHC. Hence, the periods in which 100% of players are type Y are the same for these pairs of groups. As Table 2 reveals, the earliest period for which all 12 players in the probabilistic treatment were type Y players was period 14 (group 2, part 2) and the latest such period was period 19 (group 2, part 1). On average, over both parts, all 12 players in the probabilistic treatments were Y types by period 16 of the 20 period game, which is also the theoretically expected period for all players to be type Y .

Given that preferences within each group transitioned fully from X to Y , we next ask whether players' publicly expressed actions made a similar transition. Figure 2 shows that every group converged to playing action Y by the final periods in each part of the *low* conformity treatments. Outcomes in the high conformity treatment were mixed, however, with one group (out of 6) in each of the DHC and PHC treatments never reaching the all- Y equilibrium, instead continuing to play almost exclusively action X for all 20 periods in *both* parts of the session, even after all players had switched to being Y types. Specifically, group 1 of the DHC treatment and group 2 of the PHC treatment failed to transition from playing all- X to playing all- Y . We summarize this behavior as our first finding:

Finding 1. *Most groups transition from the all- X equilibrium to the all- Y equilibrium. How-*

Group-level behavior by time, parts 1 and 2

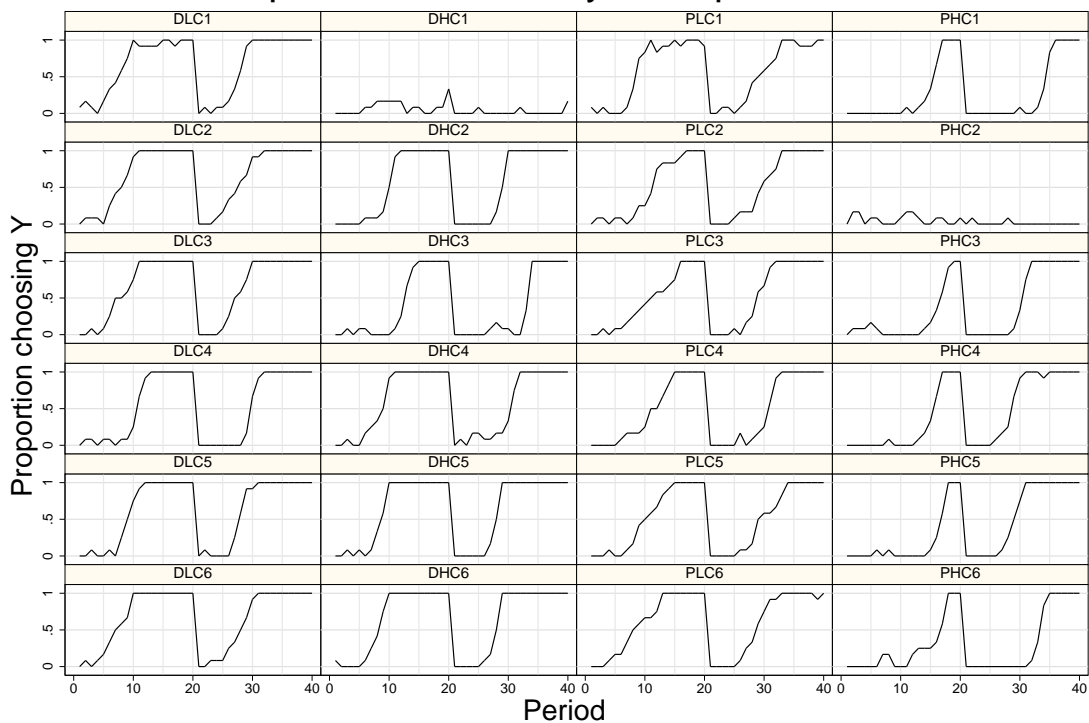


Figure 2: Proportion of subjects choosing action Y in each period of both parts of the experiment. Part 1: periods 1-20, part 2: periods 21-40.

Table 2: Speed of transition to 100% Y -types

Treatment	Group	Part 1	Part 2
PLC, PHC	1	15	16
PLC, PHC	2	19	14
PLC, PHC	3	18	15
PLC, PHC	4	18	17
PLC, PHC	5	17	16
PLC, PHC	6	16	15
PLC, PHC	Mean	17.2	15.5
DLC, DHC	All	13	13

Reported values are the first period in which a group consisted of 100% Y -types. All deterministic groups are reported in a single row, as they were guaranteed to reach 100% Y -types in period 13.

ever, some groups in the high conformity treatments never make this transition, inefficiently remaining in the all- X equilibrium.

While the failure to transition to the more efficient equilibrium was relatively uncommon in our experiment, the fact that it occurred twice (and for both parts 1 and 2) suggests that extreme cases of pluralistic ignorance are indeed possible, and can result in outcomes that are undesirable to every member of the group.

In addition to examining whether or not choice transitions occur, we also consider the nature of subjects' choices during those transitions. Here, we provide a brief, intuitive summary of subject choice behavior, with more detailed and rigorous analysis to follow. Table 3 presents treatment-level summary statistics from all periods of both parts 1 and 2. The same summary statistics, disaggregated at the group-level and divided up between parts 1 and 2 of the experiment, are reported in Tables 7 and 8 of the Appendix.

The first statistic in Table 3, the proportion of players choosing action Y , is a simple measure of the overall popularity of the newly adopted preference, Y . This proportion is, on average, higher for the low conformity treatments (DLC, PLC) as compared with the respective high conformity treatments (DHC, PHC) reflecting the predicted earlier onset and longer duration of the transition from all X to all Y in those low conformity treatments. This behavior accords with intuition that, with less pressure to conform, individuals are more

Table 3: Treatment-level Summary Statistics

	Treatment			
	DLC	DHC	PLC	PHC
$a_i = Y$	0.63 (0.48)	0.47 (0.50)	0.54 (0.50)	0.31 (0.46)
$a_i = Y \theta_i = Y$	0.90 (0.31)	0.67 (0.47)	0.87 (0.34)	0.50 (0.50)
$a_i = Y \theta_i = X$	0.07 (0.25)	0.07 (0.25)	0.07 (0.26)	0.04 (0.20)
$a_i \neq \theta_i$	0.09 (0.29)	0.24 (0.43)	0.11 (0.31)	0.31 (0.46)

Standard errors in parentheses.

willing to adopt new behaviors. The next two statistics report the proportion of players choosing action Y *conditional* on their type, either Y or X . In other words, these variables show the frequency of honesty among Y -types and dishonesty among X -types. We see that Y -type players are more likely to choose action Y in the DLC and PLC treatments as compared with the DHC and PHC treatments, which again may reflect the earlier predicted onset of the transition in the LC treatments. The final row shows that preference falsification (an action choice different from one’s type) occurs in all four of our treatments ranging from rates of approximately 10 percent in the two low conformity treatments to rates that are 2 to 3 times higher in the two high conformity treatments.

Before examining treatment differences, we first consider whether there was any change in behavior between parts 1 and 2 of the experiment. Recall that part 2 was a repetition of part 1 involving the same 12 subjects, but with new randomized draws for when each subject would transition from being type X to being type Y . Table 4 reports *differences* in behavior between the first and second parts of the experiment, showing part 2 behavior minus part 1 behavior, by treatment and across all four treatments (“pooled”) along with the results of Wilcoxon signed rank tests of the statistical significance of such differences. We observe that, for most statistics and treatments, there is no significant difference in subject behavior between parts 1 and 2. The main exceptions are for the PHC treatment, where the proportion choosing Y is marginally significantly larger in part 2 as compared with part 1 (even more significantly for type Y subjects), and in the DHC treatment where preference falsification was approximately 3 percentage points higher in part 2 as compared with part 1. Given these modest differences,

Table 4: Change in behavior between parts 1 and 2

	Treatment				
	DLC	DHC	PLC	PHC	Pooled
$a_i = Y$	0.014 (0.031)	-0.013 (0.062)	0.033 (0.027)	0.181* (0.096)	0.054** (0.053)
$a_i = Y \theta_i = Y$	0.013 (0.011)	-0.031 (0.021)	0.011 (0.025)	0.235** (0.089)	0.057 (0.031)
$a_i = Y \theta_i = X$	0.015 (0.018)	0.026 (0.014)	-0.014 (0.035)	0.074* (0.037)	0.025* (0.015)
$a_i \neq \theta_i$	-0.004 (0.007)	0.029** (0.011)	-0.010 (0.029)	-0.080 (0.051)	0.016 (0.016)

Values are group mean part 2 behavior minus group mean part 1 behavior. Standard errors in parentheses. Significance levels from two-sided Wilcoxon signed-rank tests. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

we summarize the difference between part 1 and part 2 behavior as follows:

Finding 2. *There is little evidence of learning between parts 1 and 2 of the experiment.*

Based on Finding 2, we focus the remainder of our analysis on part 2 behavior alone, after subjects have experience with the environment. Our results are very similar for part 1 behavior (results are available upon request).

We now turn to differences in behavior across treatments. Table 5 reports treatment-level differences for the same variables reported on in Table 3 but using only part 2, group-level data. We first note that there is significantly greater choice of action Y in the LC treatments than in the HC treatments, again potentially reflecting the earlier onset of the transition in the LC treatments - we will address this timing issue in more detail below. Second, we see that there is significantly more honesty, or less preference falsification, among Y -types in both low conformity treatments, relative to the respective high conformity treatments, but there is no corresponding difference for X -types. The final row of Table 5 reports preference falsification *unconditional* on type or action, and again shows that low conformity settings lead to significantly more subjects taking actions corresponding to their type. We summarize the latter findings as follows:

Finding 3. *Preference falsification occurs, and is higher in the high conformity treatments (DHC, PHC) as compared with the low conformity treatments (DLC, PLC).*

Table 5: Treatment differences

	DLC-DHC	PLC-PHC	DLC-PLC	DHC-PHC	LC-HC	D-P
$a_i = Y$	0.165** (0.054)	0.156 (0.091)	0.081** (0.023)	0.072 (0.133)	0.160** (0.067)	0.077** (0.073)
$a_i = Y \theta_i = Y$	0.246*** (0.136)	0.265** (0.137)	0.025 (0.029)	0.044 (0.191)	0.255*** (0.921)	0.034 (0.107)
$a_i = Y \theta_i = X$	-0.004 (0.031)	-0.013 (0.040)	0.008 (0.020)	-0.001 (0.046)	-0.008 (0.024)	0.003 (0.024)
$a_i \neq \theta_i$	-0.167*** (0.085)	-0.168*** (0.074)	-0.010 (0.019)	-0.010 (0.111)	-0.168*** (0.054)	-0.010 (0.064)

Observations are at the group-level, from part 2 only. Standard errors in parentheses. Significance levels from two-sided Wilcoxon signed-rank tests. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Finding 3 tells us that increased pressure for conformity discourages subjects from taking their preferred action. The differences reported in Table 5 are means across all periods, however, and thus provide only a coarse view of the relationship between action choices and underlying preferences. We next ask how popular a preference must be before it becomes accepted behavior, i.e., whether the empirical transition thresholds differ from the theoretical ones.

We first look at transitions in terms of the proportion of Y types in the population. Figures 3 and 4 show the proportion choosing action Y (vertical axis) as a function of the *proportion* of Y type players in the population (horizontal axis). Figure 3 shows the average proportion of Y choices across all six groups of each treatment relative to theoretical transition points shown as vertical bars, while Figure 4 shows the choice behavior disaggregated for each of the 24 groups individually. Note that while our transition thresholds, τ_1 and τ_2 , were originally expressed in terms of the *number* of Y types in the population, here we report these thresholds in terms of the *proportion* of the population, e.g., reporting $\tau_1 = 4/12 = 1/3$ instead of $\tau_1 = 4$. Figure 3 suggests that, on average, transitions do indeed begin earlier in the low conformity treatments as compared with the high conformity treatments. Further, the beginning of the transition in the LC treatments is approximately equal to the τ_1 threshold, while the timing of the transition for the HC treatments, appears to start, on average, a little earlier than the respective τ_1 threshold. Note further that for the LC treatments, the second threshold τ_2 is also a good indicator of when the transition to the all- Y equilibrium is complete. By contrast, the transition for the HC treatments, which is predicted to be of a shorter duration, is not complete

by the respective τ_2 threshold. The reason for the latter finding is due, in part, to there being one group (out of six) in each of the two HC treatments that never make the transition to the all-Y equilibrium in the second part of the experiment. The disaggregated group level choices shown in Figure 4 reveals such heterogeneity in outcomes across the different groups, though we observe that transitions from playing all X to playing all Y, when they occur, often start at a lower proportion of Y types in the low conformity treatments (DLC, PLC) as compared with the corresponding high conformity treatment (DHC, PLC), which is consistent with Hypothesis 4.

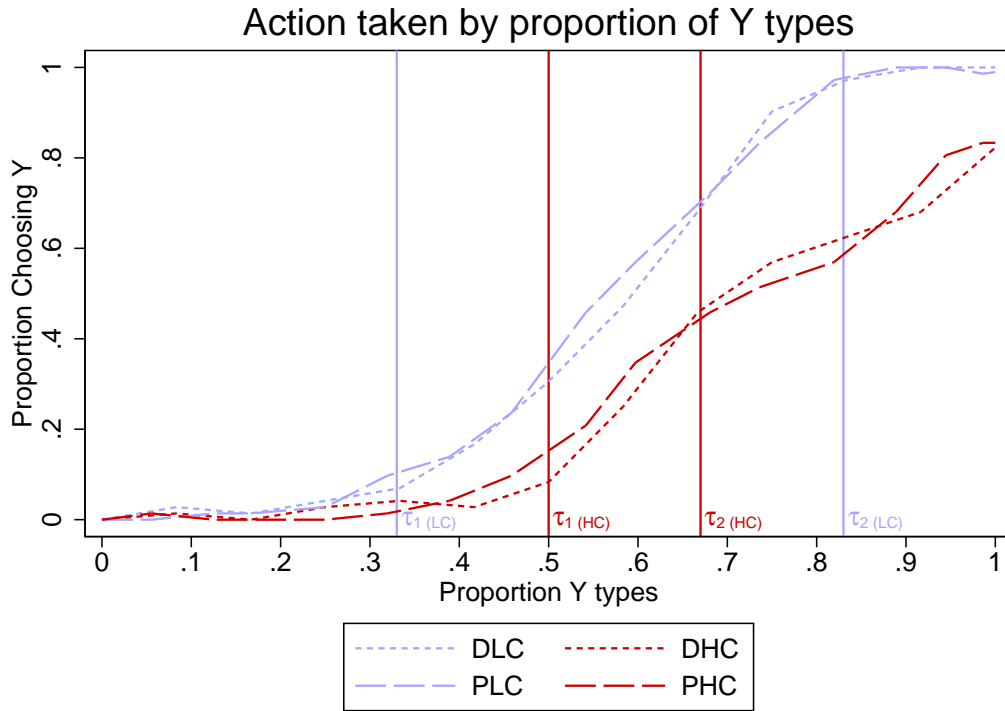


Figure 3: Average action taken by proportion of Y-types in the second part of each treatment. Vertical lines are the proportions at which transitions are predicted to begin and end.

We next consider the time it takes for transitions to occur, expressed in terms of *periods*. For the deterministic treatments, the theoretical transition periods are simply $\tau_1 + 1$ and $\tau_2 + 1$, but in the probabilistic treatments, the population is slower to transition to any given proportion of Y-types, and does so in different periods for different groups. Table 6 shows the speed with which groups transitioned between various different frequencies of playing action Y in terms of elapsed periods. Table 9 in the Appendix shows the same data, disaggregated

Group-level behavior by proportion Y types

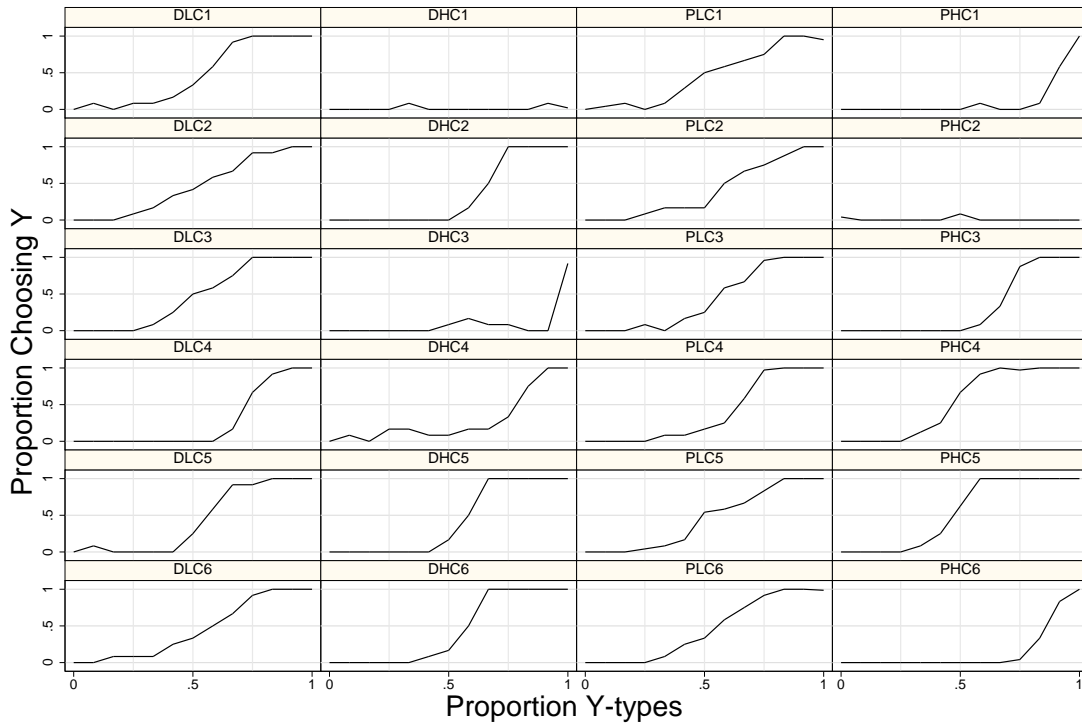


Figure 4: Proportion of subjects choosing action Y in each group as a function of the proportion of Y types in the second part of the experiment.

at the group level.

The first column of Table 6 shows how quickly (how many periods, on average) groups transitioned to having at least 25% of their members (at least 3 out of 12) playing action Y . We use a 25% threshold as a conservative indicator of when behavior has started to transition toward an eventual conversion to unanimous Y actions. Lower thresholds are difficult to interpret, as there are several instances in which one or two subjects choose Y early on, perhaps “putting their toe in the water” before switching back to playing X for several periods. Using the 25% threshold ignores these brief deviations, identifying the beginning of a more sustained transition to all Y choices. The second column of Table 6 represents the theoretical prediction for the onset of transitions, showing how quickly groups reached the initial τ_1 threshold. The final three columns of Table 6 represent different measures of the length of transitions: the number of periods needed to transition from 25% to 100% Y actions, and importantly, τ_1 to τ_2 and τ_1 to 100%, two measures of the transition phase during which we should observe both types taking their preferred actions.⁵

The significantly negative values for DLC - DHC and PLC - PHC for the first and second columns of Table 6 demonstrate that transitions begin earlier in time in the LC treatments than in the comparable HC treatments, providing support for Hypothesis 1. Note that for the transition from 0 to 25%, the difference PLC - PHC is negative but only weakly significant.

For the two groups (one DHC, one PHC) that never fully transitioned to playing all- Y , we code the period at which they achieve 25%, 100%, τ_1 , or τ_2 Y choices as period number 21, since these groups never achieve any of the four thresholds within the 20 periods of the game. Table 10 in the appendix shows transition speeds if the two non-transitioning groups are excluded. The results in Tables 6 and 10 are similar, the primary changes being that the statistical significance of the differences DLC - DHC and PLC - PHC is greater when the non-transitioning groups are included, for the obvious reason that these two HC groups did not make a transition. In particular, when we exclude the non-transitioning groups, the difference PLC - PHC is no longer weakly significant for the 0 - 25% transition.

Finding 4. *Groups in the low conformity treatments begin the transition from the all- X equilibrium to the all- Y equilibrium sooner than groups in than the equivalent high conformity treatments.*

Transitions not only start sooner with lower conformity, but also last longer. The length of

⁵The first threshold, τ_1 , is the point at which Y types are predicted to become truthful. As a result, τ_1 represents both the number of types and the number of actions we expect to see when the group begins the transition from playing X to playing Y . Likewise, once there are τ_2 Y types, all subjects are predicted to play action Y , regardless of their type.

Table 6: Speed of transitions

	0% to 25%	0% to τ_1	25% to 100%	τ_1 to τ_2	τ_1 to 100%
DLC	6.00 (0.63)	6.50 (0.56)	4.00 (0.58)	2.33 (0.49)	3.50 (0.56)
DHC	10.50 (2.05)	10.83 (2.06)	1.00 (0.26)	0.50 (0.22)	0.67 (0.21)
PLC	7.50 (0.43)	8.00 (0.45)	4.50 (0.43)	3.17 (0.60)	4.00 (0.52)
PHC	11.33 (2.01)	12.17 (1.87)	2.00 (0.45)	0.17 (0.17)	1.17 (0.31)
DLC - DHC	-4.50*** (2.14)	-4.33** (2.13)	3.00*** (0.63)	1.83*** (0.54)	2.83*** (0.60)
PLC - PHC	-3.83* (2.06)	-4.17** (1.92)	2.50*** (0.62)	3.00*** (0.62)	2.83*** (0.60)
DLC - PLC	-1.50** (0.76)	-1.50** (0.72)	-0.50 (0.72)	-0.83 (0.78)	-0.50 (0.76)
DHC - PHC	-0.83 (2.87)	-1.33 (2.78)	-1.00** (0.52)	0.33 (0.28)	-0.50 (0.37)

Values are mean number of periods elapsed between the given thresholds in the second part of the experiment. Thresholds are proportions of group members playing action Y . Non-transitioning groups are coded as reaching each threshold in period 21, i.e., beyond the 20th period of the second part. Standard errors in parentheses. Significance levels for treatment differences (DLC-DHC and PLC - PHC) are from one-sided Mann-Whitney U tests. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

the transition phase is longer in the LC treatments relative to the comparable HC treatments as evidenced by the significantly positive value for the differences DLC - DHC and PLC - PHC over the 3rd, 4th and 5th columns in Table 6. The differences in these three measures of the duration of the transition to the all- X equilibrium provide support for Hypothesis 2. Figure 3 also illustrates that the speed of transition is greater in the HC treatments as compared with the LC treatments.

Finding 5. *Groups in the low conformity treatments spend longer in the transition phase than groups in the equivalent high conformity treatments.*

Having shown that pressures for conformity significantly influence both the timing and duration of transitions, we next consider whether the structure of the progression of types (deterministic or probabilistic) similarly influence transitions, in line with Hypotheses 3 and 4. The differences DLC - PLC and DHC - PHC show that all transitions occur sooner and are shorter in the deterministic treatments than in the probabilistic treatments, though with mixed levels of significance. In the low conformity treatments, the transition occurs sooner in the deterministic treatment than in the probabilistic treatment, though the *length* of the transition is not significantly different. In the high conformity treatments, there is no significant difference in the onset of the transition, however duration of the transition is significantly shorter in the deterministic treatment than in the probabilistic treatment. Collectively, these differences provide moderate support for Hypothesis 3.

Finally, we again examine transition speed in terms of number of Y types rather than number of periods. Focusing on action choices in relation to the *proportion* of Y -types holds constant the popularity of a preference across groups and treatments. Recall that Hypothesis 4 states that, conditional on the number of Y -types, there should be no difference between the deterministic and probabilistic versions of the same conformity treatment (LC) or (HC). Table 11 in the Appendix reports the same transition threshold differences as in Table 6 but reporting transitions in terms of the number of Y -types instead of the number of periods. Table 11 reveals virtually no difference between DLC and PLC or DHC and PHC. The lack of significant differences in transition speed by the number of Y -types supports Hypothesis 4.

Finding 6. *Transitions occur sooner in the deterministic treatments than in probabilistic treatments when measured in terms of number of periods, but no sooner when measured in terms of number of Y -types.*

Finding 6 says that, while the timing of transitions is slowed in the probabilistic treatments, that difference appears to be a product of the smaller number of Y -types in any given period, and not the uncertainty about the number of Y -types in the population.

Beyond the timing and duration of transitions, there are also differences in the level of disagreement that occurs during the transition phase. The earlier and slower transitions in the low conformity environments allow for a larger share of the population to publicly join in the minority position during those transitions. Figure 5 shows the fraction of the population taking the minority action as a function of the proportion of Y -types in the population. The size of the minority can be thought of as a *measure of disagreement* within each group. When the minority is small, most people are taking the same action, whereas when the percentage taking the minority position approaches 50% the group is at maximal disagreement. As Figure 5 reveals, the average size of the minority is significantly larger in the LC treatments than in the HC treatments (9.5% versus 4.9%, $p < 0.001$, Mann-Whitney). Subjects in the LC treatments are more likely to take their preferred action, leading to higher levels of disagreement for longer periods of time. The tradeoff for having relatively low levels of disagreement as in the HC treatments can be seen in Figure 6, which shows the fraction of subjects taking the opposite action from their type as a function of the proportion of Y -types in the population. Behavior in all treatments is similar when the population is primarily X -types, but falsification in the HC treatments exceeds that in the LC treatments once the number of Y -types exceeds approximately one-third of the population.

Summarizing, on the one hand, higher conformity environments lead to more sudden transitions between states, which can mean shorter periods of disagreement. On the other hand, higher conformity environments are more likely to get stuck in inefficient equilibria involving extreme pluralistic ignorance, where universally desirable change does not occur. There is therefore a tension between the likelihood of having a long, painful transition and the likelihood of being stuck in a bad equilibrium. Low conformity environments are more likely to complete transitions, but those transitions are marked by longer stretches of disagreement. High conformity environments are more likely to become stuck in bad equilibria, but have more rapid transitions if they do occur.

6 Conclusion

Social change is a complicated process. The evolution of privately held preferences interacts with societal pressures to conform to perceived social norms in ways that can affect the speed of social change. We have provided a simple game theoretic model of this process and provided some experimental evidence in support of our model's main predictions. Consistent with the predictions of our model, we find evidence for public preference falsification in the laboratory, and that preference falsification affects the onset and speed of social change. Greater incentives for conformity (stronger social pressures) lead to public behavior that is slower in tracking

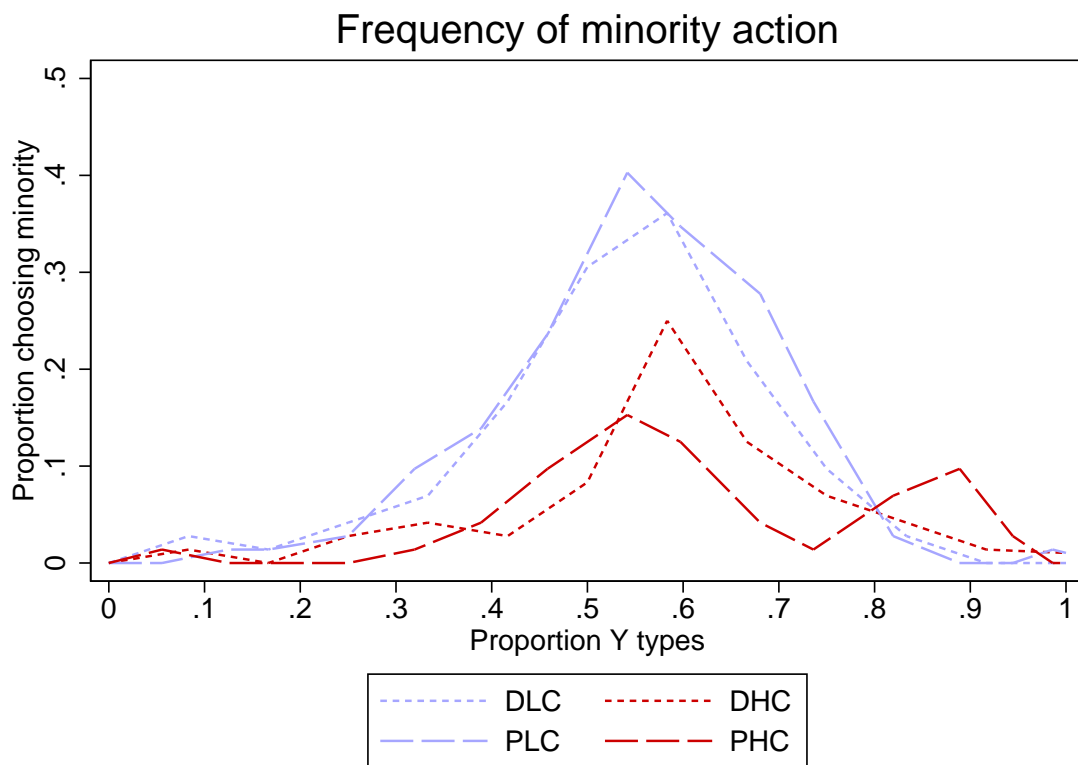


Figure 5: Frequency of minority action choices by proportion of Y-types in the second part of each treatment.

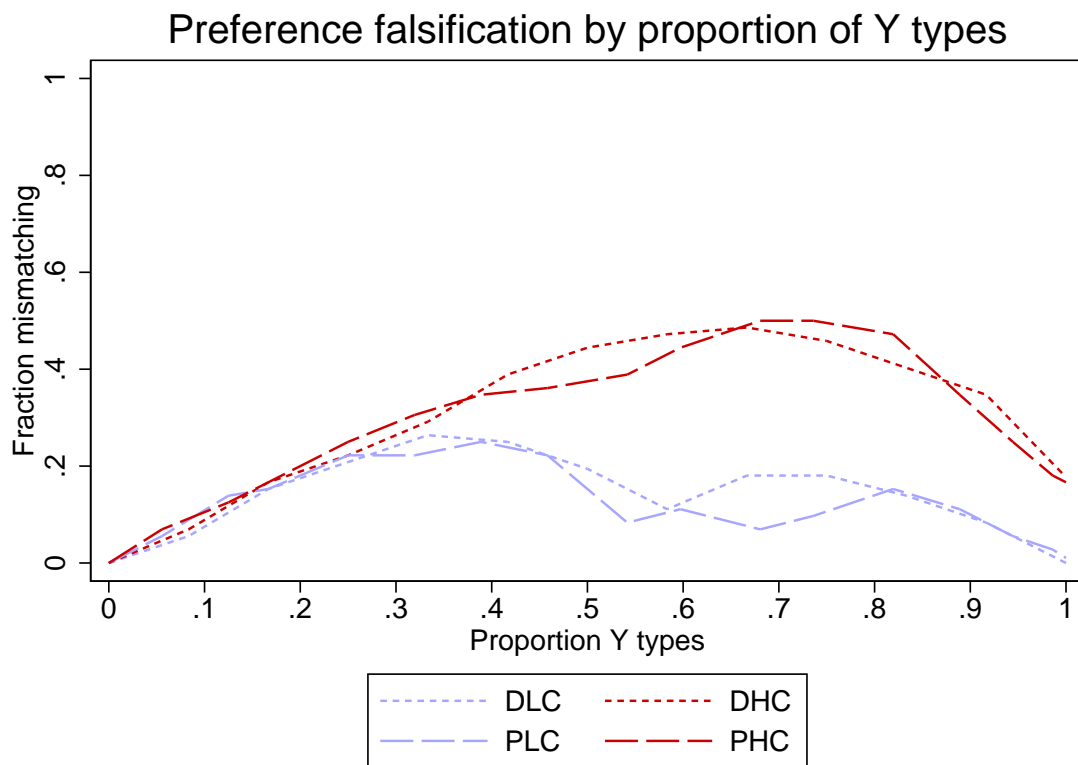


Figure 6: Frequency of preference falsification by proportion of Y-types in the second part of each treatment.

the evolution of privately held preferences. The slower transition results in relatively large numbers of people “living lies,” though with the benefit of little disagreement between groups; if a collective change in behavior occurs, it does so quickly. By contrast, when incentives for conformity are low, we see public behavior more closely tracking private preferences, involving a longer transition phase accompanied by relatively high levels of disagreement. In our low conformity treatment, the change from the all X equilibrium to the all Y equilibrium always occurred, but in the high conformity treatment there are instances where a transition never occurred, despite the new behavior being universally preferred to the old.

We have only considered two different processes for the transition of private preferences. It would be of interest to consider other preference evolution processes or mechanisms that might work to speed up or slow down the transitions between equilibria. For instance, one could change the knowledge individuals have about the process by which types change, or the feedback available about what actions were taken by others in each period. Alternatively, one could allow for a more “lumpy” evolution of preferences, e.g., where more than one player’s preferences can transition each period, or a non-monotonic evolution, where the preferences of some players waffled back and forth for some length of time between the two alternatives. Finally, it would, of course, be useful to develop a more structural model that endogenized the evolution of preferences, as opposed to our reduced form, exogenous transition process. We leave these important extensions to future research.

Appendix

Table 7: Group-level Choices (part 1 only)

	Prop. choosing $a_i = Y$	Prop. choosing $a_i = Y \theta_i = Y$	Prop. choosing $a_i = Y \theta_i = X$	Prop. choosing $a_i \neq \theta_i$
DLC1	0.66	0.90	0.15	0.12
DLC2	0.65	0.93	0.06	0.07
DLC3	0.64	0.93	0.04	0.06
DLC4	0.52	0.77	0.00	0.16
DLC5	0.58	0.85	0.03	0.11
DLC6	0.67	0.96	0.08	0.05
DLC Mean	0.62	0.89	0.06	0.10
DHC1	0.08	0.12	0.00	0.59
DHC2	0.54	0.79	0.03	0.15
DHC3	0.41	0.61	0.00	0.27
DHC4	0.61	0.86	0.09	0.12
DHC5	0.61	0.85	0.10	0.13
DHC6	0.63	0.88	0.10	0.11
DHC Mean	0.48	0.69	0.05	0.23
PLC1	0.58	0.83	0.15	0.17
PLC2	0.48	0.81	0.14	0.17
PLC3	0.48	0.90	0.03	0.06
PLC4	0.47	0.88	0.05	0.08
PLC5	0.51	0.86	0.08	0.11
PLC6	0.60	0.92	0.05	0.07
PLC Mean	0.52	0.87	0.08	0.11
PHC1	0.27	0.42	0.00	0.38
PHC2	0.06	0.11	0.02	0.45
PHC3	0.23	0.44	0.01	0.29
PHC4	0.27	0.53	0.01	0.24
PHC5	0.20	0.37	0.00	0.35
PHC6	0.26	0.41	0.00	0.38
PHC Mean	0.22	0.38	0.01	0.35

Table 8: Group-level Choices (part 2 only)

	Prop. choosing $a_i = Y$	Prop. choosing $a_i = Y \theta_i = Y$	Prop. choosing $a_i = Y \theta_i = X$	Prop. choosing $a_i \neq \theta_i$
DLC1	0.66	0.93	0.12	0.09
DLC2	0.65	0.94	0.05	0.05
DLC3	0.66	0.93	0.09	0.08
DLC4	0.54	0.78	0.03	0.15
DLC5	0.64	0.90	0.10	0.10
DLC6	0.65	0.93	0.06	0.07
DLC Mean	0.63	0.90	0.08	0.09
DHC1	0.02	0.03	0.00	0.66
DHC2	0.58	0.82	0.10	0.16
DHC3	0.39	0.57	0.00	0.29
DHC4	0.55	0.78	0.08	0.18
DHC5	0.63	0.87	0.14	0.13
DHC6	0.64	0.87	0.15	0.14
DHC Mean	0.47	0.66	0.08	0.26
PLC1	0.55	0.90	0.05	0.08
PLC2	0.55	0.90	0.01	0.07
PLC3	0.58	0.88	0.10	0.11
PLC4	0.51	0.80	0.11	0.16
PLC5	0.53	0.88	0.05	0.09
PLC6	0.59	0.91	0.09	0.09
PLC Mean	0.55	0.88	0.07	0.10
PHC1	0.32	0.52	0.01	0.29
PHC2	0.01	0.01	0.01	0.61
PHC3	0.51	0.77	0.08	0.17
PHC4	0.60	0.89	0.21	0.15
PHC5	0.58	0.88	0.18	0.14
PHC6	0.36	0.59	0.00	0.25
PHC Mean	0.40	0.61	0.08	0.27

Table 9: Group-level speed of transitions

	0% to 25%	0% to τ_1	25% to 100%	τ_1 to τ_2	τ_1 to 100%
DLC1	6	6	3	2	3
DLC2	5	5	6	4	6
DLC3	5	6	4	3	3
DLC4	9	9	2	1	2
DLC5	6	7	4	1	3
DLC6	5	6	5	3	4
DLC Mean	6.00	6.50	4.00	2.33	3.50
DHC1	Never	Never	Never	Never	Never
DHC2	8	8	1	1	1
DHC3	12	13	1	0	0
DHC4	9	10	2	0	1
DHC5	7	7	1	1	1
DHC6	7	7	1	1	1
DHC Mean	10.50 (8.60)	10.83 (9.00)	1.00 (1.20)	0.50 (0.60)	0.67 (0.80)
PLC1	7	7	5	5	5
PLC2	8	8	4	4	4
PLC3	7	8	4	2	3
PLC4	9	10	3	1	2
PLC5	8	8	5	4	5
PLC6	6	7	6	3	5
PLC Mean	7.50	8.00	4.50	3.17	4.00
PHC1	13	14	2	0	1
PHC2	Never	Never	Never	Never	Never
PHC3	9	10	2	0	1
PHC4	7	8	3	0	2
PHC5	7	8	3	1	2
PHC6	12	13	2	0	1
PHC Mean	11.33 (9.60)	12.17 (10.60)	2.00 (2.40)	0.17 (0.20)	1.17 (1.40)

Values are number of periods elapsed between the given thresholds in the second part of the experiment. Non-transitioning groups are coded as reaching each threshold after 20 periods. Means with non-transitioning groups excluded are reported in parentheses.

Table 10: Speed of transitions, excluding non-transitioning groups

	0% to 25%	0% to τ_1	25% to 100%	τ_1 to τ_2	τ_1 to 100%
DLC	6.00 (0.63)	6.50 (0.56)	4.00 (0.58)	2.33 (0.49)	3.50 (0.56)
DHC	8.60 (0.93)	9.00 (1.14)	1.20 (0.20)	0.60 (0.25)	0.80 (0.20)
PLC	7.50 (0.43)	8.00 (0.45)	4.50 (0.43)	3.17 (0.60)	4.00 (0.52)
PHC	9.60 (1.25)	10.60 (1.25)	2.40 (0.25)	0.20 (0.20)	1.40 (0.25)
DLC - DHC	-2.60** (1.09)	-2.50** (1.20)	2.80*** (0.66)	1.73** (0.59)	2.70*** (0.65)
PLC - PHC	-2.10 (1.22)	-2.60** (1.23)	2.10*** (0.52)	2.97*** (0.69)	2.60*** (0.61)
DLC - PLC	-1.50** (0.76)	-1.50** (0.72)	-0.50 (0.72)	-0.83 (0.78)	-0.50 (0.76)
DHC - PHC	-1.00 (1.56)	-1.60 (1.69)	-1.20*** (0.32)	0.40 (0.32)	-0.60** (0.32)

Values are the mean number of periods elapsed between the given thresholds in the second part of the experiment. Thresholds are proportions of group members playing action Y . Standard errors in parentheses. Significance levels for treatment differences (DLC-DHC and PLC - PHC) are from one-sided Mann-Whitney U tests. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 11: Speed of transitions by number of Y -types

	0% to 25%	0% to τ_1	25% to 100%	τ_1 to τ_2	τ_1 to 100%
DLC	6.00 (0.63)	6.50 (0.56)	4.00 (0.58)	2.33 (0.49)	3.50 (0.56)
DHC	9.33 (1.05)	9.50 (1.06)	0.83 (0.31)	0.50 (0.22)	0.67 (0.21)
PLC	6.00 (0.37)	6.67 (0.33)	4.17 (0.31)	2.83 (0.48)	3.50 (0.34)
PHC	8.67 (1.33)	9.33 (1.17)	1.67 (0.42)	0.33 (0.21)	1.00 (0.26)
DLC - DHC	-3.33** (1.23)	-3.00** (1.20)	3.17*** (0.65)	1.83*** (0.54)	2.83*** (0.60)
PLC - PHC	-2.67 (1.38)	-2.67 (1.22)	2.50*** (0.45)	2.50*** (0.52)	2.50*** (0.43)
DLC - PLC	0.00 (0.73)	-0.17 (0.65)	-0.17 (0.65)	-0.50 (0.69)	0.00 (0.66)
DHC - PHC	0.67 (1.70)	0.17 (1.58)	-0.83 (0.52)	0.17 (0.31)	-0.33 (0.33)

Values are the mean proportion of subjects who are Y -types when each threshold is met. Thresholds are the proportion of group members playing action Y . Non-transitioning groups are coded as reaching each threshold with 13 Y -types. Standard errors in parentheses. Significance levels for treatment differences are from one-sided Mann-Whitney U tests for LC versus HC comparisons and two-sided Mann-Whitney U tests for D versus P treatments. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

References

- Allport, Floyd H. 1924. *Social Psychology*. Boston: Houghton-Mifflin.
- Anderson, Hans Christian. 1838. *Fairy Tales Told for Children. First Collection*. Copenhagen: C. A. Reitzel.
- Anderson, Lisa R., and Charles A. Holt. 1997. "Information Cascades in the Laboratory." *American Economic Review*: 847-862.
- Andreoni, James, Nikos Nikiforakis and Simon Siegenthaler. 2017 "Social Change and the Conformity Trap." working paper.
- Asch, Solomon E., 1956. *Studies of Independence and Conformity. A Minority of One Against a Unanimous Majority*. *Psychological Monographs*, 70(9), 1-70.
- Charness, Gary, Luca Rigotti, and Aldo Rustichini. 2007. "Individual Behavior and Group Membership." *American Economic Review*, 97(4): 1340-1352.
- Chwe, Michael Suk-Young. 2001. *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton: Princeton University Press.
- Cooper, Russell, Douglas V. DeJong, Robert Forsythe and Thomas W. Ross. 1989. "Communication in the Battle of the Sexes Game: Some Experimental Results." *The RAND Journal of Economics*, 20(4): 568-587.
- Cooper, Russell, Douglas V. DeJong, Robert Forsythe and Thomas W. Ross. 1993. "Forward Induction in the Battle-of-the-Sexes Games." *The American Economic Review*, 83(5): 1303-1316
- Crawford, Vincent P., Uri Gneezy, and Yuval Rottenstreich. 2008. "The Power of Focal Points Is Limited: Even Minute Payoff Asymmetry May Yield Large Coordination Failures." *American Economic Review*, 98(4): 1443-58.
- Dana, Jason, Roberto A. Weber and Jason Xi Kuang. 2007. "Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory* 33(1): 57-80.
- DellaVigna, Stefano, John A. List and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics*, 127(1): 1-56.

- Devetag, Giovanna and Andreas Ortmann, 2007. "When and Why? A Critical Survey on Coordination Failure in the Laboratory." *Experimental Economics* 10(3): 331-344
- Goeree, Jacob K., and Leeat Yariv. 2015. "Conformity in the Lab." *Journal of the Economic Science Association*, 1(1): 15-28.
- Hung, Angela A., and Charles R. Plott. 2001. "Information Cascades: Replication and an Extension to Majority Rule and Conformity-Rewarding Institutions." *American Economic Review*: 1508-1520.
- Kuran, Timur. 1995. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Cambridge, MA: Harvard University Press.
- Michaeli, Moti, and Daniel Spiro. 2015. "Norm Conformity Across Societies." *Journal of Public Economics*, 132, 51-65.
- Michaeli, Moti, and Daniel Spiro. 2017. "From Peer Pressure to Biased Norms." *American Economic Journal: Microeconomics* 9(1): 152-216.
- Moreno, Bernardo and María del Pino Ramos-Sosa. 2017. "Conformity in Voting." *Social Choice and Welfare* 48(3), 519-543.
- Moscovici, Serge. 1985. *Social Influence and Conformity*. In: G. Lindzey and E. Aronson (Eds.), *The Handbook of Social Psychology*, 3rd ed., Vol. 2, New York: Random House, pp. 347-412.
- O'Gorman, Hubert J. (1975). "Pluralistic Ignorance and White Estimates of White Support for Racial Segregation." *Public Opinion Quarterly*, 39, 313-330.
- Ochs, Jack. 1995, "Coordination Problems" in J.H. Kagel and A.E. Roth (Eds.), *The Handbook of Experimental Economics*. Princeton: Princeton University Press, pp. 195-251.
- Prentice, Deborah A. and Dale T. Miller. 1993. "Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm." *Journal of Personality and Social Psychology* 64(2), 243-56.
- Sherman, Steven J., Clark C. Presson, Laurie Chassin, Eric Corty and Richard Olshavsky. 1983. "The False Consensus Effect in Estimates of Smoking Prevalence: Underlying Mechanisms." *Personality and Social Psychology Bulletin*, 9(2), 197-207.
- Turner, John C. 1991. *Social influence*. Pacific Grove, CA: Brooks/Cole.