Social Conformity Under Evolving Private Preferences*

John Duffy† Jonathan Lafky‡

September 1, 2020

Abstract

We propose a model of how social norms change in response to the evolution of privately held preferences. Our aim is to rationalize the tendency for individuals who hold minority preferences to take actions favored by the majority. We do this using a game involving a tension between a desire to act according to one's underlying preferences and a desire to conform to the majority opinion. In an experimental setting, we find that even after a majority of the population shares what was previously an unpopular minority opinion, members of the new majority are slow to change their behavior. The timing and speed with which behavior transitions to match new, majority-held opinions depends on the size of the reward for conformity. When the rewards for conformity are low, the transition is gradual, with considerable periods of costly public disagreement. When the rewards for conformity are high, transitions are slow to start but conclude rapidly once they begin.

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. The experimental protocol was approved by the UC Irvine Institutional Review Board, HS# 2015-2082. Tyler Boston provided expert research assistance. We thank audiences at several conferences and workshops for helpful comments and suggestions. An earlier version of this paper was circulated under the title "Living a Lie: Theory and Evidence on Public Preference Falsification."

†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 own private preferences over outcomes. 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 to fit in with colleagues on a Friday night. Other examples are more impactful, such as a person expressing public opinions about issues like drug legalization, same sex marriage, the #MeToo movement or free speech on campus 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 the social norm. 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, one way to frame the discussion is to adopt Kuran's (1987, 1995) terminology of "preference falsification" which refers to situations where individuals express a position that is opposite to their own privately held position because of a desire to conform to the behavior of the majority. The extent of preference falsification can vary; in the most extreme cases, a majority of the population expresses positions that are only truly held by a minority of the society.[1] Importantly, the preferences of the majority may be imperfectly known, and thus perceptions of majority opinion may be incorrect, a situation that social psychologists refer to as "pluralistic ignorance" (Allport 1924). Pluralistic ignorance arises when individuals mistakenly believe that they hold minority beliefs. Game theoretically, pluralistic ignorance implies an absence of common knowledge (Chwe 2001). 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 drinking on college campuses (Prentice and Miller 1993) and tolerance of smoking by others (Sherman et al. 1983), despite majorities opposed to such activities.

We show how the preference falsification phenomenon can be captured using a dynamic,

[1]The 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.

n-player, non-cooperative coordination game and we examine how subjects play this game in a laboratory experiment. In our model, players are induced to hold private preferences about some binary issue and must make a binary, public choice about the same issue, e.g. the expression of an opinion to others. Following Bernheim (1994), Goeree and Yariv (2015), and Bernheim and Exley (2015), each player's utility depends on the extent to which their public choice conforms with the choices of the other $n-1$ players – the coordination game aspect of our game– as well as the extent to which their choice is consistent with their own private preference or "type"; the tension between social conformity and being true to one's own type can lead to public preference falsification. Specifically, a player's utility from expressing a certain public preference is increasing in the number of others expressing that same preference, but expressing a preference that is at odds with one's own private preference (type) brings disutility to the player. We model these gains and losses from conformity explicitly in terms of the utility or payoff that agents receive. Further, in our framework, all players' preferences are known to evolve over time, which allows for interesting dynamics. Initially, all players' private preferences are the same. Over time, each player's private preferences change from the original preference to the opposite preference in either a deterministic or stochastic manner so that over a sufficiently long period of time all n players' private preferences will have switched from a preference for one action to the other. For example, initially all players might hold the private preference that cigarette smoking is acceptable behavior in public spaces. Over time, due to (say) evidence that smoking causes lung cancer, players individually switch their private preference to opposing smoking in public spaces. This evolution of private preferences over time is a key feature of our approach that enables us to identify whether, and to what extent, expressed, public preferences may depart from these private preferences.

Preference falsification can be seen as both a cause and effect of coordination failure. 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 social norm (or equilibrium) even after a majority have switched to privately preferring an alternate outcome. Coordination failures in turn lead to falsification when an old social norm persists even after preferences have evolved in favor of a different norm. 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. While this paper is not the first to study preference falsification or pluralistic ignorance - see the literature review for some relevant key references– what we add is the use of game theory and experimental economic methods to model and empirically document the preference falsification phenomenon.

Kuran (1995) suggests that, like Darwin, we should look to historical, natural experiments

for supportive evidence of public preference falsification. However, 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). The main problem with studying preference falsification in the field is that it may not be possible to know both individuals' true private preferences and majority opinion at any 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 the more natural settings of social conformity that we seek to model. By inducing players to hold preferences for conformity and by being neutral about the decision-making context, we can clearly evaluate how accurately aggregate group choices reflect individuals' privately held preferences versus social conformity to majority opinion. 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.

The coordination game we use is similar to a repeated n-player Battle of the Sexes coordination game, but with private preferences gradually changing over time from all players preferring one outcome to all players eventually favoring the opposite outcome. Our game has many dynamic equilibria, ranging from extreme preference falsification where behavior never transitions from the initial outcome, to less extreme preference falsification with gradual transition from one outcome to the other, as well as a most efficient equilibrium, in which the transition from one outcome to the other is immediate and unanimous once the majority's private preferences favor the new outcome. We use this rich environment to ask several questions. First, how often does preference falsification occur, and how long can it be sustained after a majority of group members have abandoned their old, privately-held preference? In other words, which of the many dynamic equilibrium paths is the most empirically plausible? We find in our experiment that preference falsification does occur, but that it is seldom extreme and seems to follow simple regularities. Most of our groups *do* transition from one social norm to the other, as private preferences evolve. However, it is neither abrupt nor unanimous, as in the most efficient equilibrium. Second, we ask whether variations in the payoffs to conformity play any role in such transitions. We find that a larger incentive to conform results in greater preference falsification and a delayed onset of the transition to the new social norm, but with the advantage of transitions that are over quickly once they begin. By contrast, smaller incentives to conform result in transitions that are quick to start, but drag on over longer period of time, resulting in considerably greater disagreement among players as to which option to coordinate upon. When pressures to conform are high we do find a small number of severe cases of preference falsification where the initially efficient stage game

equilibrium is sustained even after *all* members of the group have adopted private preferences that would make coordination on the other outcome the more efficient choice. Finally, we ask whether greater uncertainty about the majority position, i.e., the potential for pluralistic ignorance, affects the timing of the transition to the new equilibrium by considering both deterministic and stochastic processes for the evolution of private preferences. We find that greater uncertainty does delay the period in which a transition occurs, but this is mainly owing to the delay that our stochastic transition process induces in the evolution of private preferences. Taking that delay into account, the timing of the transition depends primarily on when the majority's preferences switch from preferring one outcome over the other.

2 Related literature

The notion of preference falsification and its social consequences was first elaborated upon by Kuran (1987, 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. Key differences between this work and the present paper are that we are using both game theory and experimental economics methods to study the problem, we avoid deception, and we allow preferences to evolve over time.

Indeed, the problem we study can be viewed as a coordination game with heterogeneous preferences, as in the Battle of the Sexes game, of which there exist several experimental studies, including Cooper et al. (1989, 1993), Charness et al. (2007), and Crawford et al. (2008) among others.[2] 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. More closely related to our study, Jeitschko and Taylor (2001) present a dynamic, n-player, Pareto ranked coordination game in which uncertainty about the state of nature, payoff relevant complementarities in action choices and private information about outcomes can result in an inevitable "coordination avalanche" from an efficient to a less efficient equilibrium, even when fundamentals would favor sustained adoption of the efficient equilibrium by all. Guarino et al. (2006) however, provide experimental evidence against this coordination avalanche outcome as subjects in their experiment seem to ignore the possibility that other

[2]Ochs (1995) and Devetag and Ortmann (2007) provide surveys of this experimental coordination game literature.

subjects are having experiences different from their own, which they term a solipsism bias. While we also have uncertainty about the state of the world and complementarities in payoffs, in our game, aggregate outcomes and payoffs are always public information so that solipsism biases are minimized. Further, in our game, the Pareto optimal equilibrium changes over time as preferences evolve, so that our game is more like an n-player Battle of the Sexes game than the n-player Stag Hunt game of Jeitschko and Taylor (2001).

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 this rational herding phenomenon was first provided by Anderson and Holt (1997). Despite some similarities, there are important differences between our environment and the information herding model. First, players in our game make decisions simultaneously, rather than sequentially, so that they cannot condition their choice in the current period on the choices made earlier by others in that same period; they can only condition on choices made by others in prior periods. Second, preferences are private information and are known to be evolving over time so that the state of the world is also changing, unlike in the standard social learning model. Consequently, agents have to be more forward-looking in their decision-making and cannot rely upon learning from past outcomes alone.

Also closely related to our work is a literature that addresses "innate" preferences for conformity. Bernheim (1994) provides a theoretical model of conformity in which individuals are concerned with their own status, which in turn is determined by other people's beliefs about the individual's own private preferences or "predispositions." A person with sufficient concern for their own status will conform to a single socially acceptable behavior in order to signal to others that they have "good" preferences, even when their true preferences might be for a different action. Bernheim and Exley (2015) use a laboratory setting to demonstrate the existence of innate preference-based tastes for conformity, in addition to any external factors such as social sanctions or social learning. Similarly, Andreoni and Bernheim (2009) show that a concern for being perceived by others as being fair strongly influences behavior toward the "50-50 norm" of equitable division in the dictator game. Akerlof and Kranton (2000) develop a model of identity in which both internal and external pressures encourage conformity. People exhibit behavior that is typical of their identity group so as to maintain their own perceived membership within that group. Individuals also sanction groupmates who deviate from acceptable behavior, as "mavericks" harm the identity of everyone in the group.

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 and independent research, Andreoni et al. (2019) and Smerdon et al. (2020) also study social change in the laboratory with evolving preferences. Andreoni et al. (2019) focus on probabilistically evolving preferences, while we compare and contrast environments with deterministic and probabilistic change in preferences. Also differently from our study, in the Andreoni et al. design players are matched, payoffs are assessed and information is revealed *pairwise*, whereas payoffs in our setting depend on *group decisions* (i.e., we employ n-player matching), since we have in mind that aggregate information about social preferences is publicly revealed. As in our study, Andreoni et al. find 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. Nevertheless, we view our study as complementary to theirs, especially concerning knowledge about the population-wide distribution of types or actions.

Smerdon et al. (2020) implement a version of Brock and Durlauf's (2001) discrete choice model with social interactions (which involves a somewhat different payoff function from our model) to investigate the persistence of "bad social norms" by which they mean persistence in play of an equilibrium that becomes socially inefficient over time. In their design, subjects choose between two "doors" but have incomplete information about the option preferred by others. Subjects get payoffs based on their own private value and a social value that is increasing (decreasing) in the number of others making the same (opposite choice) choice as their own. Unknown to subjects, one option is preferred by a large margin in the first half of the experiment while the other option becomes more valuable in the second half of the experiment. They find that bad social norms are more likely to persist in environments where subjects are uncertain of the preferences of others and the strength of social value is strong. We obtain similar results in a less complex setting, e.g., we have only two player types (as opposed to a continuum) and the process by which preferences evolve in our setting is completely known to participants. Still, we obtain qualitatively similar findings which is

re-assuring regarding the robustness of the phenomenon of the persistence of bad social norms.

3 Theory

In this section we outline the theoretical model that we will implement and test in our experiment. 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. The first environment is the simplest possible setting, in which 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 per person and are both permanent and irreversible, 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 Model

The game consists of n players, each of whom makes an action choice in each of the $t = 1, 2, ..., T \geq n$ 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 an 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 $H > L > 0$ and k_{a_i} is the total number of players who choose action a_i in that period 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 privately 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

[3]It 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).

players and $c_j^Y = \min\{j-1,\ n\}$ Y-type players. We chose this stark 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 avoided. In the next section we consider a more natural stochastic transition process, which has the effect of creating the possibility of 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.

3.2 Equilibria

Taken as a whole, our dynamic coordination game admits many equilibria; a complete characterization is beyond the scope of this paper. Some of these equilibria involve no *disagreement*, by which we mean that all players take the same action within each period. For instance, all players choosing action X in all periods or all players choosing action Y in all periods are both equilibrium possibilities. Similarly, there exist equilibria where players alternate every period between all choosing X or all choosing Y. A further equilibrium without any disagreement, and the most efficient in total payoff terms, involves all players choosing action X until the point at which half of the players have switched from type X to type Y, at which point all players immediately switch over to playing action Y for the duration of the game. There are also equilibria that exhibit disagreement, in which players choose different actions, with X-type players choosing X while Y-types choose Y.

Because of the wide array of possible equilibria in the full dynamic game, we focus our analysis on the feasibility of certain types of *stage game* equilibria as the number of Y-types in the population evolves over time. Specifically, we focus on the progression from early periods in which a disagreement stage-game equilibrium is not feasible, to intermediate periods which do support disagreement equilibria, to later periods in which a disagreement equilibrium is again infeasible.[4]

Note that, in any stage-game equilibrium, all subjects of the same type must take the same action as one another. If an equilibrium existed in which subjects of the same type took different actions, some of those subjects would prefer X to Y while others preferred Y

[4]For a dynamic model of switching from one strict Nash equilibria to another by boundedly rational players in a two-strategy coordination game, see Lyu (2020).

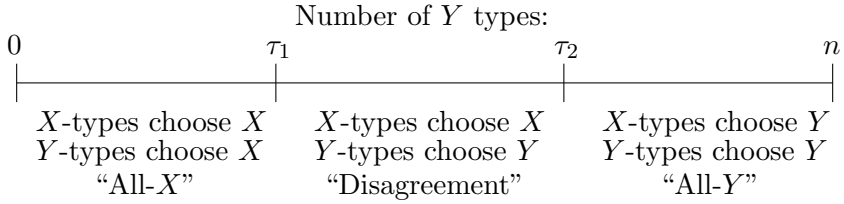

Figure 1: Feasible range of disagreement equilibria

to X, which is impossible since the payoff to taking either action is increasing in the number of players choosing that action.

To see when a disagreement equilibrium is possible, consider a scenario in which each player chooses an action equal to their type. Each Y-type prefers to play Y if it yields a higher payoff than switching to $a_i = X$, or equivalently, when the payoff from choosing Y, or $H \cdot \frac{c^Y}{n}$, is greater than the payoff from switching to playing X instead, or $L \cdot \frac{c^X+1}{n}$. Since $c^X = n - c^Y$, this condition is equivalent to $c^Y > \frac{(n+1)L}{H+L}$. In other words, $c^Y > \frac{(n+1)L}{H+L}$ 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, a disagreement equilibrium ceases to be feasible when each X-type player is no longer willing to choose an observable action $a_i = X$, even if all other X-types chose X. This is true when $L \cdot \frac{c^Y+1}{n} > H \cdot \frac{c^X}{n}$, or equivalently $c^Y > \frac{nH-L}{H+L}$. Once $c^Y > \frac{nH-L}{H+L}$, an equilibrium with disagreement is no longer sustainable, resulting once again in all players taking the same action.

Let $\tau_1(H, L, n)$ denote the smallest number of Y-types at which disagreement becomes feasible, i.e., $\tau_1(H, L, n) = \left\lceil \frac{(n+1)L}{H+L} \right\rceil$, and let $\tau_2(H, L, n)$ denote the smallest number of Y-types at which disagreement is no longer possible, i.e., $\tau_2(H, L, n) = \left\lceil \frac{nH-L}{H+L} \right\rceil$. 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") 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 dynamics are illustrated in Figure 1.

As Figure 1 makes clear, there are two kinds of players who might rationally choose to "live a lie" i.e., misstate their type - those whose private type switches from type X to type Y in the first τ_1 periods but who announce X during those same periods and those who privately remain type X in the periods starting from period $\tau_2 + 1$ onward and who 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 illustrate how the viability of disagreement equilibria can change over time, we present two numerical examples that are also used in our experiment. Let $n = 12$, $H = 9$ and $L = 3$. A disagreement equilibrium becomes possible once $c^Y > \frac{(n+1)L}{H+L} = \frac{(12+1)\cdot 3}{9+3} = 3.25$, meaning that $\tau_1 = 4$, which occurs in period $t = 5$. At this point, a disagreement equilibrium would have $k_Y = 4$, and $k_X = 8$, as all X-type players are still willing to choose $a_i = X$. Disagreement remains feasible until $c^Y > \frac{nH-L}{H+L} = \frac{9\cdot 12-3}{9+3} = 8.75$, meaning that $\tau_2 = 9$, which occurs in period $t = 10$. From period 10 onward, all players are predicted to choose the same action.

Alternatively, suppose that $n = 12$, $H = 9$ and $L = 6.43$. In this case the ratio between the H and L payoffs is smaller, so the pressure to conform to the majority opinion will be greater than in the previous example. With this new parameterization, all players are predicted to take the same action until $c^Y > \frac{(n+1)L}{H+L} = \frac{(12+1)\cdot 6.43}{9+6.43} = 5.42$, meaning that $\tau_1 = 6$, which occurs in period $t = 7$. Both Y and X-types can act in accordance with their type until the second threshold is crossed, when $c^Y > \frac{nH-L}{H+L} = \frac{12\cdot 9-6.43}{9+6.43} = 6.58$, meaning that $\tau_2 = 7$, which occurs in period $t = 8$. In other words, beginning in period 8, the pressure to conform is great enough that disagreement once again becomes infeasible.

This numerical example illustrates two dynamic patterns that we will look for in our experimental data. First, when the ratio H/L is large so that the incentive to conform with the choices of others is low, the minimal number of Y-types needed for the *onset* of a transition, τ_1, is lower and thus a transition (if one occurs) can happen *earlier* in time as compared with the case where the ratio H/L is smaller and the incentives to conform are greater. Second, when the ratio H/L is large, if a transition occurs, the period of "disagreement," as defined by the difference $\tau_2 - \tau_1$, will be *longer* than when the ratio H/L is smaller; in the latter case transitions are relatively more abrupt. Of course, it also remains a possibility that there will never be a transition (no disagreement) or there will be alternating periods of agreement on different social norms as well; equilibrium selection remains an empirical question. However, we speculate that, as the payoffs associated with the two agreement equilibria (all-X and all-Y) change, so too will the behavior of the human subjects in our experiment.

3.3 Probabilistic Model

In many scenarios involving preference falsification, there is uncertainty as to the proportion of the population that prefers each action, i.e., there is the potential for 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 $t > 1$ there is a commonly known constant probability, p, that exactly one player, randomly selected among the set of all X-type players as of period $t-1$, switches from being an X-type player to being a Y-type and remains a Y-type in all subsequent periods, $t+1, \ldots T$. Under this probabilistic model, the expected number of Y-type players in period $t = 1$ is 0, the expected number in period $t = 2$ is p, and the expected number in period t is $p(t-1)$, until all Y-types have switched to X-types. Recognizing that the number of Y-type players can never exceed n, the expected count of Y-types in period t is $E[c_t^Y] = \min\{p(t-1),\ n\}$ and likewise, the expected count of X-types is given by $E[c_t^X] = \max\{n - p(t-1),\ 0\}$. In addition to decreasing the expected number of Y types in each period relative to the deterministic case, this change in structure also has the effect of creating uncertainty about the actual number of Y types, and, in later rounds, what the majority type is.

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 choice of X by all n players to a choice of Y by all n players, if it occurs, should come later in time in the probabilistic case as compared with the deterministic case; how much later in time depends upon 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 is $L = \$6.43$. These parameterizations were chosen in order to test the comparative statics of the theory by varying the feasibility of the onset, duration and completion of the transition from the all-X stage game equilibrium to the all-Y stage game equilibrium. As discussed earlier in the numerical examples of Section 3.1, in the LC treatments, $\tau_1 = 4$ and $\tau_2 = 9$. By contrast, in the HC treatments, $\tau_1 = 6$, and $\tau_2 = 7$, so in the HC treatments, the transition phase, if it occurs, 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 $t = 2, 3, \ldots$ 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 17 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 has occurred by the final, 20th period of the game. Using a normal approximation to the binomial distribution, the probability of such a complete transition, in which all 12 players are Y-types by the 20th period is approximately 93 percent.[5]

In the probabilistic treatments, each group $g = 1, 2, ...6$ of treatment PLC experiences an independent, randomly determined sequence of type progressions using $p = 0.75$ in each of parts 1 and 2. We then use those *same* random sequences of type progressions for parts 1 and 2 for one matched group $g = 1, 2, ...6$ 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

[5]The game involves 20 periods of play. In 19 of these periods, the probability of successful switch from X-type to Y-type is 0.75. The probability of having 12 such successes in 19 trials is approximated by $\Phi\left(\frac{12-19\cdot 0.75}{\sqrt{0.25\cdot 0.75\cdot 19}}\right) \approx 0.93$.

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
H	\$9.00	\$9.00	\$9.00	\$9.00
L	\$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]	9 [10]	7 [8]	9 [13]	7 [11]

[a]As 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 disagreement Nash equilibrium in which all X-types play X and all Y-types play Y, i.e., there exists an equilibrium with full revelation of types. When there are at least 9 Y-types in the population (which occurs in period 10), this fully-revealing 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) indicated between brackets in Table 1, 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 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 players for both parts. We chose to keep subjects in the

same matching group of 12 in the second part so as to create conditions most favorable to examination of any role played by experience, which we examine later in the paper.

At the start of each session, subjects received instructions for the first part of the session only. These instructions were read aloud in an effort to make the information public knowledge. The instructions avoided any 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 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. Subjects then completed the first part of the experiment consisting of 20 periods of decisions. Following completion of the first part, subjects received instructions for the second part. While the second part was a repeat of the first part, subjects did not know this fact in advance. Following completion of the second part, subjects were paid their earnings from two randomly chosen periods, one drawn from each of the two 20-period parts.

Parts 1 and 2 began identically, with all $n = 12$ subjects starting out as 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.

At the beginning of each period, prior to making any choices, each subject was privately informed about their type for the period 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.

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.[6] 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. 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, nine in the LC and seven in the HC treatments, X-types now prefer to switch away from a disagreement equilibrium to collectively choosing action Y. Thus, as noted earlier, in the the LC treatments there exists a disagreement equilibrium where agents choose the action corresponding to their own type so long as the expected number of Y-types is between four and eight, inclusive. The disagreement equilibrium in the HC treatments is shorter-lived, becoming feasible when the expected number of Y-types is six, and no longer being feasible when the expected number is seven.

With the foregoing analysis in mind, 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 smaller 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 can "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.*

Hypothesis 2 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 disagreement will last longer in the low conformity treatments than in the high conformity treatments. Indeed, as Table 1 reveals, the number of Y-types needed before the disagreement equilibrium ceases to be feasible, τ_2, is *higher*, and thus comes later in time, in the low conformity treatments as

[6] In other words, we are considering whether there exists a group of players for which each player within that group is individually better off if the group collectively chooses Y instead of X.

compared with the comparable high conformity treatments. Intuitively, 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 begins earlier and is of shorter duration 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 where they are delayed by a factor of $1/p$. Thus, Hypothesis 3 is effectively a robustness 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. Our theory predicts that τ_1 and τ_2 are the same between DLC and PLC and between DHC and PHC (see again Table 1), however it is possible that the process by which preferences evolve may affect behavior, separate from the number of Y-types in any given period. In particular, in the probabilistic treatment, subjects are uncertain as to exactly how many other players there are of each type, leading to the possibility of pluralistic ignorance. The lack of common knowledge concerning the number of each player type may result in further delay between changes in types and changes in actions.

Hypothesis 4. *The number of Y-types at which the disagreement phase begins and ends is larger in the probabilistic treatments than in the equivalent deterministic treatments. (PLC versus DLC and PHC versus DHC).*

Intuitively, Hypothesis 4 says that, holding pressure for conformity constant, the tendency to disagree may depend not only on the popularity of an opinion, but also on the process by which that 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.

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.

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, each subject's type had switched from X to Y before the final, 20th period in both parts 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, it took 16.4 periods before all 12 players in the probabilistic treatments were Y types, just shy of the theoretical prediction of 17 periods.

Given that preferences within each group transitioned fully from X to Y, we next ask whether players' 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 similar, however, 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 (a total of 40 periods) 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:

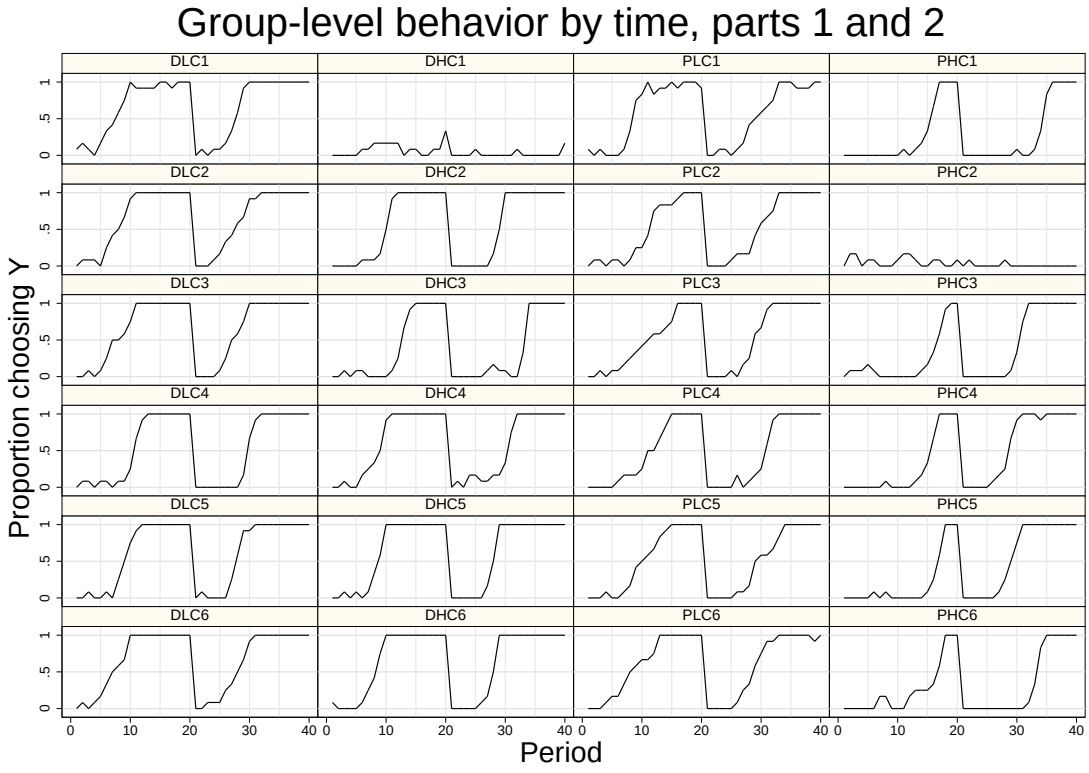

Figure 2: Proportion of subjects choosing action Y in each period of both parts of the experiment. Part 1: periods 1-20, part 2: periods 21-40.

Finding 1. *Most groups transition from the all-X equilibrium to the all-Y equilibrium. Only one group in each of the high conformity treatments never transitions, 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) shows that these bad outcomes are indeed possible, despite being clearly undesirable for every individual in the group.

Table 3: Treatment-level Summary Statistics

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

Standard errors in parentheses.

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 8 and 9 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 the intuition that, with less pressure to conform, individuals are more willing to adopt new behaviors. The next two statistics report the proportion of players choosing action Y *conditional* on their type being either Y or X. In other words, these variables show the frequency with which Y-types take their preferred action and X-types take their less-preferred action, i.e., preference falsification. 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

Table 4: Change in behavior between parts 1 and 2

	Treatment				
	DLC	DHC	PLC	PHC	Pooled
$a_i = Y$	0.014	-0.013	0.033	0.181*	0.054**
	(0.031)	(0.062)	(0.027)	(0.096)	(0.053)
$a_i = Y \mid \theta_i = Y$	0.013	-0.031	0.011	0.235**	0.057
	(0.011)	(0.021)	(0.025)	(0.089)	(0.031)
$a_i = Y \mid \theta_i = X$	0.015	0.026	-0.014	0.074*	0.025*
	(0.018)	(0.014)	(0.035)	(0.037)	(0.015)
$a_i \neq \theta_i$	-0.004	0.029**	-0.010	-0.080	0.016
	(0.007)	(0.011)	(0.029)	(0.051)	(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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

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

Finding 2. *There is little difference in behavior 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

Table 5: Treatment differences

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

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

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 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).*

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, are expressed in terms of the *number* of Y types in the population, it is clearer at times to discuss these thresholds in terms of the *proportion* of the population, which we denote below as $\tau_i^p = \tau_i/n$. For example, $\tau_1^p = \tau_1/n = 4/12 = 1/3$ when $\tau_1 = 4$.

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^p threshold, while the timing of the transition for the HC treatments, appears to start, on average, a little earlier than the respective τ_1^p threshold. Note further that for the LC treatments, the second threshold τ_2^p 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^p 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.[7] The disaggregated group level choices, as shown in Figure 4, reveal 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, PHC).

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 for the probabilistic treatments, the population is slower to transition to any given number 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 10 in the Appendix shows the same data, disaggregated 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

[7] Figure 8 in the appendix shows aggregate behavior similar to Figure 3, but restricted only to those groups that successfully transitioned to the all-Y equilibrium.

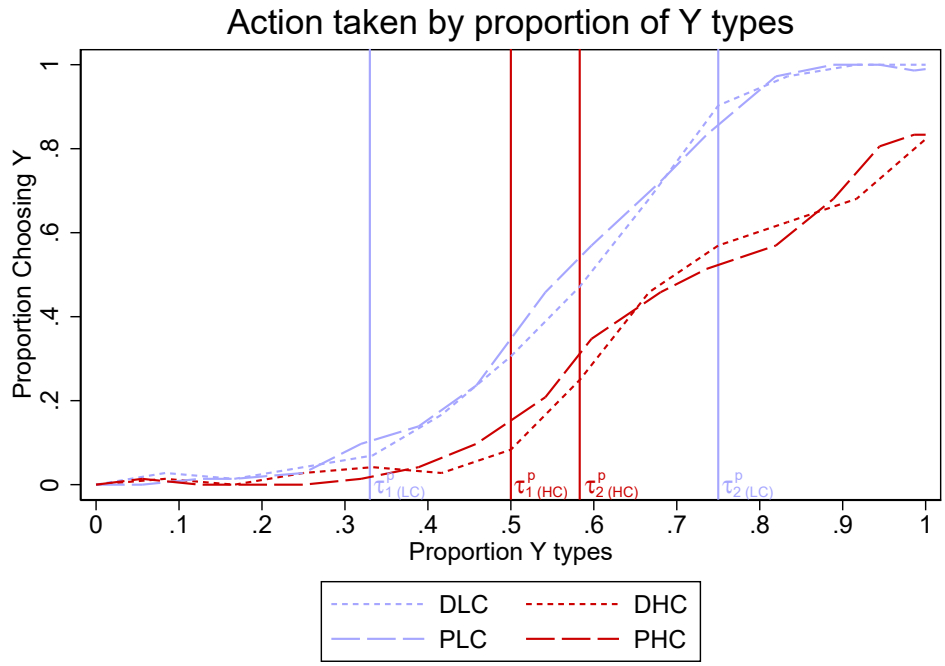

Figure 3: Average action taken by proportion of Y-types in the second part of each treatment. Vertical lines are the proportions at which transitions are predicted to begin and end.

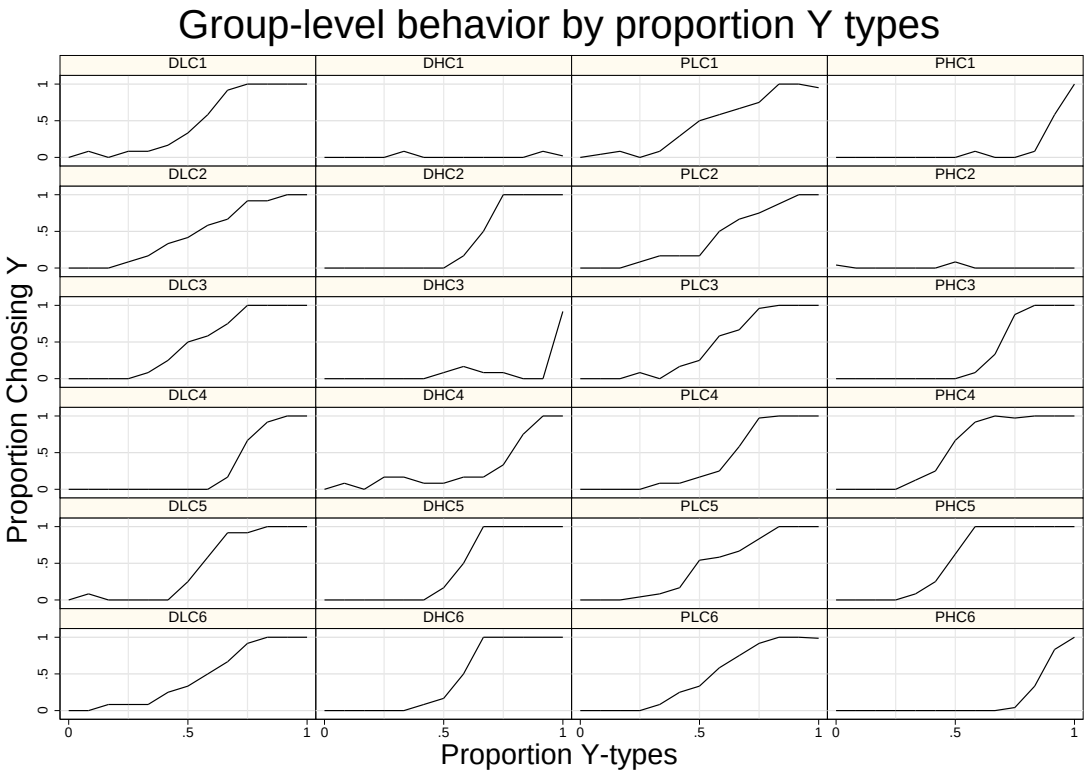

Figure 4: Proportion of subjects choosing action Y in each group as a function of the proportion of Y types in the second part of the experiment.

Table 6: Speed of transitions

	0% to 25%	0% to τ_1^p	25% to 100%	τ_1^p to τ_2^p	τ_1^p to 100%
DLC	6.00	6.50	4.00	2.17	3.50
	(0.63)	(0.56)	(0.58)	(0.48)	(0.56)
DHC	10.50	10.83	1.00	0.50	0.67
	(2.05)	(2.06)	(0.26)	(0.22)	(0.21)
PLC	7.50	8.00	4.50	2.67	4.00
	(0.43)	(0.45)	(0.43)	(0.49)	(0.52)
PHC	11.33	12.17	2.00	0.17	1.17
	(2.01)	(1.87)	(0.45)	(0.17)	(0.31)
DLC - DHC	-4.50***	-4.33**	3.00***	1.67***	2.83***
	(2.14)	(2.13)	(0.63)	(0.53)	(0.60)
PLC - PHC	-3.83*	-4.17**	2.50***	2.50***	2.83***
	(2.06)	(1.92)	(0.62)	(0.52)	(0.60)
DLC - PLC	-1.50**	-1.50**	-0.50	-0.50	-0.50
	(0.76)	(0.72)	(0.72)	(0.69)	(0.76)
DHC - PHC	-0.83	-1.33	-1.00**	0.33	-0.50
	(2.87)	(2.78)	(0.52)	(0.28)	(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 final period. Standard errors in parentheses. Significance levels for treatment differences (DLC-DHC and PLC - PHC) are from one-sided Mann-Whitney U tests. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

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^p 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^p to τ_2^p and τ_1^p to 100%, two measures of the transition phase during which we should observe both types taking their preferred actions.[8]

[8]The first threshold, τ_1^p, is the point at which Y types are predicted to begin taking their preferred action. As a result, τ_1^p 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.

The significantly negative values for DLC - DHC and PLC - PHC in 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^p, or τ_2^p Y choices as period number 21, since these groups never achieve any of the four thresholds within the 20 periods of the game. Table 11 in the appendix shows transition speeds if the two non-transitioning groups are excluded. The results in Tables 6 and 11 are similar; the primary change 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 the equivalent high conformity treatments.*

When incentives to conform are lower, transitions not only start sooner, but also last longer. The length of 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 the 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 the number of Y-types rather than the 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, conditional on the number of Y-types, we expect there to be no difference in the speed of transition between the deterministic and probabilistic versions of the same conformity treatment (LC) or (HC). Table 12 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 12 reveals virtually no difference between DLC and PLC or DHC and PHC. Hypothesis 4 allowed for the possibility that the probabilistic environment resulted in slower transitions, even holding the number of Y-types constant. We do not find such an effect, with no significant differences in transition speed by the number of Y-types, meaning that we cannot reject the null of equality between the deterministic and probabilistic treatments.

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 is not due to the greater uncertainty about the number of Y-types in the population. Finding 6 also suggests that the passage of time itself has little direct influence on subject choices, with subjects instead conditioning their behavior on the (expected) proportion of Y types at each point in time.[9]

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 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.[10] The size of the minority can be thought of as a *measure of disagreement* within each group. When the minority

[9]This behavior is worthy of further study, for example by comparing settings with various proportions of types that do not vary over time, allowing for the possibility that subjects are employing Markov-style reasoning to decide when to switch behavior.

[10]Figure 5 may at first appear inconsistent with Figure 3, such as when Figure 3 shows the average frequency of Y as being approximately 50%, while Figure 5 shows the minority size to be only approximately 10%. To see that these figures are, in fact, consistent with one another, consider a hypothetical with one group that

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.

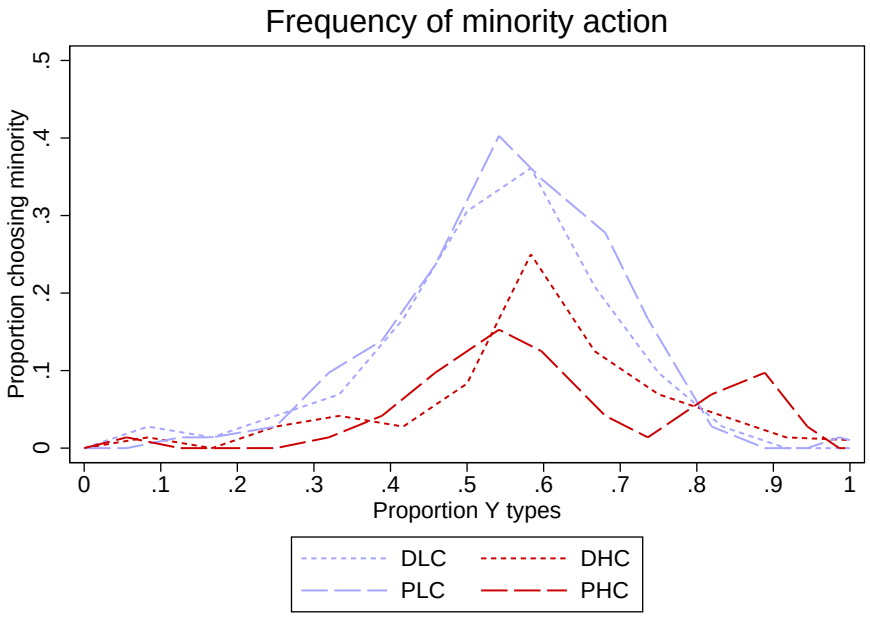

Figure 5: Frequency of minority action choices by proportion of Y-types in the second part of each treatment.

Thus far, our data analysis has focused on aggregate differences across treatments and sessions. We next turn to exploring behavior at the *individual* subject level. Specifically, we focus on determinants of an individual's decision to switch their action choice from X to Y. We first consider the impact of the timing of the one-time preference change on the decision

played 90% Y, and a second group that played 90% X. The average frequency of Y across both groups would be 50%, while the average minority size would only be 10%.

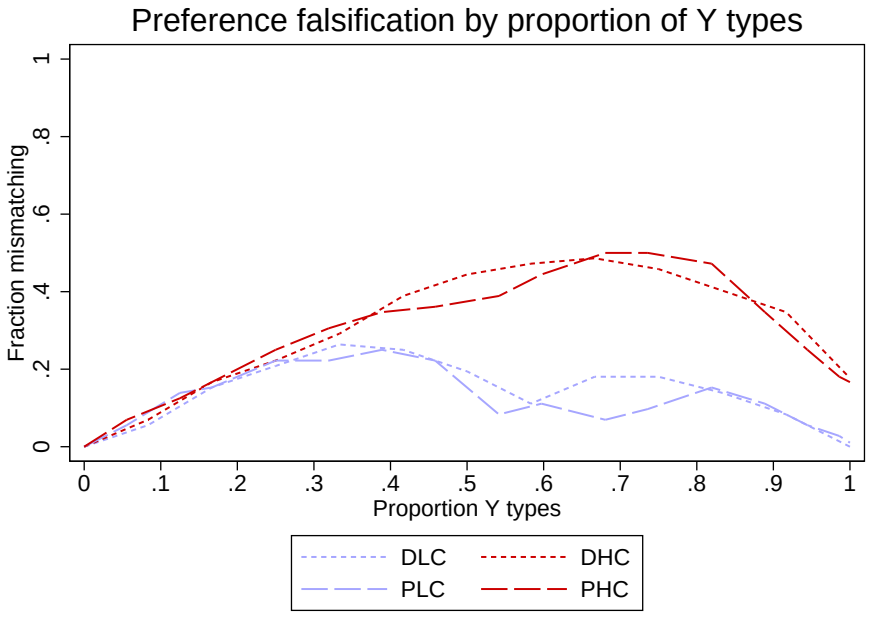

Figure 6: Frequency of preference falsification by proportion of Y-types in the second part of each treatment.

to switch action choices. Recall from our discussion of Figure 1 that there are two kinds of players who might choose to "live a lie" i.e., mismatch their type. There are players whose type changes from X to Y in the first τ_1 periods, who are "ahead-of-their-time" in terms of type, but who might not want to reveal their new preference so as to conform to existing social norms. There are also "old-fashioned" players whose true type changes late in the game, sometime after period τ_2, but who can nevertheless "see the writing on the wall" and choose to switch over to the new social norm prior to privately supporting it. Figure 7 provides clear evidence for both types, along with a considerable number of "truth-telling" subjects as well. In this figure, the period in which a subject's type permanently changes from X to Y is indicated on the horizontal axis, while the vertical axis measures the number of periods elapsed from the type change period until the subject had consistently switched their public action choice from X to Y. Subjects who delayed revealing their new Y type have a positive value for this lag, while those who switched actions prior to their type changing have a negative value for this lag; a zero lag indicates subjects whose behavior switched simultaneously with their type. As we speculated, ahead-of-their-time subjects who become Y-types in the early rounds of the game tend to delay choosing action Y, while old-fashioned subjects, whose type changes only later in the game, find it better to begin playing Y prior to their type actually switching.

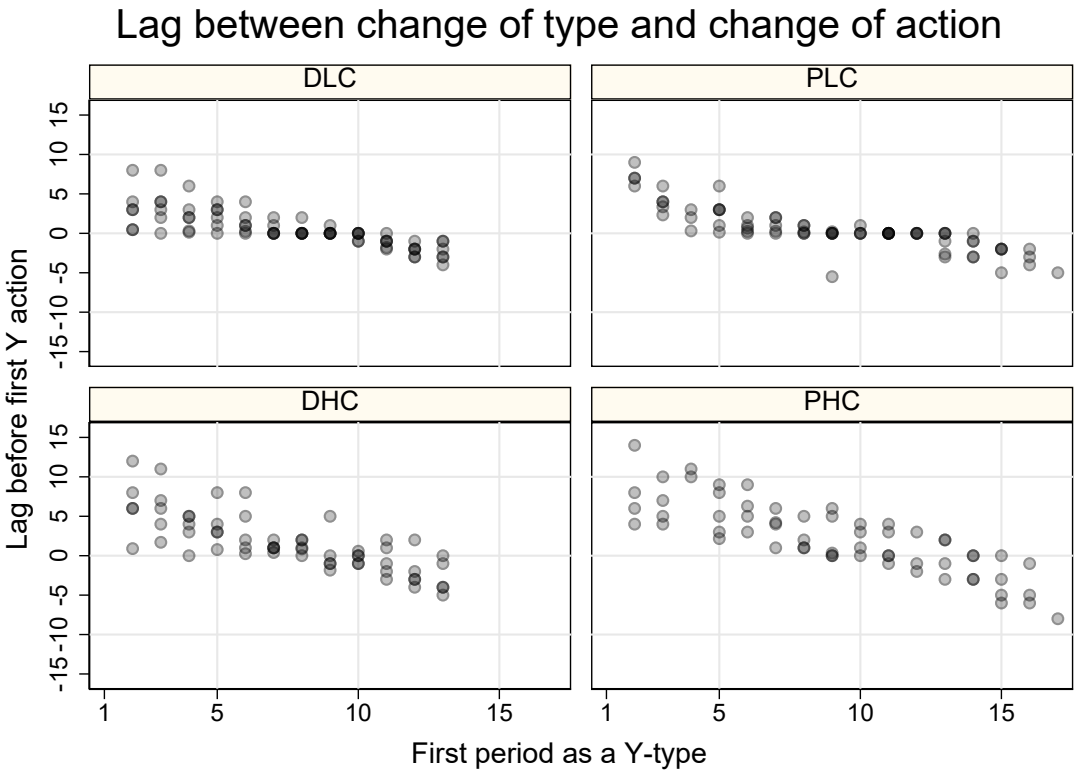

Figure 7: The x-axis shows the period in which a subject switched to being a Y-type. The y-axis shows the number of periods elapsed from the time a subject became a Y-type to the first period that began a sequence of playing Y at least twice in a row. Behavior is very similar if we instead require only a single choice of Y. Note that negative values of the lag indicate subjects who played action Y prior to them becoming a Y-type, and darker circles indicate a higher density of observations.

Notice further that the transition lag is steeper in the HC treatments as compared with the LC treatments and the transition is completed faster in the D treatments as compared with the P treatments, confirming our earlier findings using aggregate data.

We extend our examination of individual behavior by performing a regression analysis of subjects' decision to switch actions. Table 7 reports estimates from panel, probit-model regressions of subjects' switching behavior. The dependent variable in these regressions is an indicator variable that takes the value 1 if subject i in period t switched from action $a_{i,t-1} = X$ in the previous period to $a_{i,t} = Y$ in the current period, and takes the value 0 otherwise. Explanatory variables include a dummy variable $\Delta\theta_{i,t}$ indicating the period in which the subject's type changed, which is equal to 1 if subject i was an X-type in the previous period ($\theta_{i,t-1} = X$) and is a Y-type in the current period ($\theta_{i,t} = Y$), and is set to 0 otherwise. We also include a measure of how many more subjects chose action Y in the previous period $(t-1)$ than in the period before that $(t-2)$, which we denote by Δk_Y. The latter measure captures the speed with which other subjects are switching their behavior to the new norm. In addition, the regression includes linear and quadratic terms for the period number, t, to examine time trends in switching behavior.

As Table 7 reveals, all four explanatory variables are statistically significant in explaining switching behavior in the pooled estimation (using all data) and, with one exception (Δk_Y in the PLC treatment), for each of the four treatments separately as well. The change in a subject's own type is comparatively more influential in the LC treatments, while changes in the number of others choosing Y has a comparatively larger effect in the HC treatments. Finally, regarding the time trend, we see that switching increases with the passage of time, t, but at a diminishing rate, as indicated by the negative coefficient on t^2. This pattern is consistent with the idea that most switching takes place in the middle periods, with little change occurring in the earlier or later periods. Note that our period variables are also picking up the effect of the expected number of Y types at each point in time; indeed, our results are very similar if we use the expected number of Y types in place of the period number, t.

Summarizing, we find that in all treatments a subject's decision to switch from action X to action Y depends substantially on when that subject's own type changes; the change in own type has the largest marginal impact on the likelihood of switching, as shown in Table 7. However, pressures to conform and the period of the game (or the expected number of Y types) also matter for subjects' switching behavior. We find that subjects are sensitive to changes in the behavior of those around them, meaning that subjects may delay or accelerate their switch from choosing X to choosing Y as revealed in Figure 7. Our analysis of individual subject behavior largely confirms but also enriches our main findings using the aggregate data.

Table 7: Determinants of switching action from $a_{i,t-1} = X$ to $a_{i,t} = Y$

	Pooled	DLC	DHC	PLC	PHC
$\Delta\theta_{i,t}$	0.205***	0.222***	0.123*	0.329***	0.0649***
	(0.0301)	(0.0378)	(0.0639)	(0.0512)	(0.0121)
Δk_Y	0.0163***	0.0131***	0.0143***	-0.000818	0.0218***
	(0.00275)	(0.00508)	(0.00226)	(0.00633)	(0.00259)
t	0.0511***	0.0737***	0.0580***	0.0898***	0.0650***
	(0.00809)	(0.00661)	(0.0104)	(0.0173)	(0.0159)
t^2	-0.00274***	-0.00478***	-0.00314***	-0.00475***	-0.00294***
	(0.000481)	(0.000481)	(0.000735)	(0.000854)	(0.000599)
N	4752	1296	1080	1296	1080

Coefficients are marginal effects following panel probit regressions. The dependent variable is a dummy for whether the subject chose $a_{i,t} = Y$ this period and $a_{i,t-1} = X$ in the previous period. Each of the groups from the DHC and PHC treatments that did not transition to playing Y are not included, as they exhibited almost no variation in behavior. Data from the final two rounds is also dropped to remove two outliers who briefly switched between playing X and Y in the final two periods (see column 3 of Figure 2). Standard errors are clustered at the group level. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

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, or whether social change takes place at all. We have provided a simple game theoretic model in which we can observe this process. As our model admits many equilibria involving various degrees of preference falsification, we implemented our model in the laboratory to observe the dynamics of preference falsification in practice. While we do find experimental evidence of preference falsification, it tends to be neither as extreme nor as minimal as our model allows and the extent of preference falsification seems to follow simple regularities. While most of our groups transition from one social norm to the other as private preferences evolve, the timing of that transition and its duration depend on model parameters in intuitive ways. Greater incentives for conformity (stronger social pressures) lead to public behavior that is slower in tracking the evolution of privately held preferences. The delayed transition results in relatively large numbers of people "living lies," though with the benefit of little disagreement within groups; when a collective change in behavior finally occurs, it does so quickly. Strong incentives to conform also have the potential to completely prevent social change from ever occurring, though we see only a few instances of this in our experiment. These empirical findings suggest that there may be a tension between the length of transition and the probability that a transition occurs at all. By contrast, when incentives for conformity are low, we see behavior that more closely tracks private preferences, resulting in a longer transition phase accompanied by relatively high levels of disagreement.

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 that individuals have about the process by which types change, or the feedback available about what actions were taken by others in each period, or allow communication as to what actions players intend to choose. 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. Such modifications could help to isolate the mechanisms that cause groups to transition, as our design is largely silent on the process by which individual subjects chose between their options. Our model describes when disagreement equilibria are *feasible*, but it does not preclude a wide range of other possible behavior, such as remaining at the original norm indefinitely, or immediately switching to the new norm, even before it develops widespread support. Future studies should help to explain

how individuals address the equilibrium selection problem inherent in this setting. 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.

References

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

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, 87 (5): 847-862.

Andreoni, James, Nikos Nikiforakis and Simon Siegenthaler. 2019 "Social Change and the Conformity Trap." Working paper.

Andreoni, James, and B. Douglas Bernheim. 2009. "Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects." Econometrica, 77(5): 1607-1636.

Asch, Solomon E., 1956. Studies of Independence and Conformity. A Minority of One Against a Unanimous Majority. Psychological Monographs, 70(9): 1-70.

Bernheim, B. Douglas 1994. "A Theory of Conformity." Journal of Political Economy, 102(5): 841-877.

Bernheim, B. Douglas, and Christine Exley. 2015. "Understanding Conformity: an Experimental Investigation." Harvard Business School NOM Unit Working Paper No. 1

Brock, William A., and Steven N. Durlauf. 2001. "Discrete Choice with Social Interactions." Review of Economic Studies, 68(2): 235-260.

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.

Devetag, Giovanna and Andreas Ortmann, 2007. "When and Why? A Critical Survey on Coordination Failure in the Laboratory." Experimental Economics 10(3): 331-344

Guarino, Antonio., Steffen Huck, and Thomas D. Jeitschko, 2006. "Averting economic collapse and the solipsism bias." Games and Economic Behavior, 57(2), 264-285.

Jeitschko, Thomas D., and Curtis R. Taylor, 2001. "Local discouragement and global collapse: a theory of coordination avalanches." American Economic Review, 91(1), 208-224.

Kuran, Timur. 1987. "Preference Falsification, Policy Continuity and Collective Conservatism," The Economic Journal, 97(387): 642-665.

Kuran, Timur. 1995. Private Truths, Public Lies: The Social Consequences of Preference Falsification. Cambridge, MA: Harvard University Press.

Lyu, Jianxun. 2020. "Heterogeneity in Cognition and Equilibrium Selection in Coordination Games," Working paper, University of Edinburgh.

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.

Smerdon, David, Theo Offerman and Uri Gneezy. 2020. " 'Everybody's Doing It' On the Persistence of Bad Social Norms." Experimental Economics 23, 392-420.

Turner, John C. 1991. Social influence. Pacific Grove, CA: Brooks/Cole.

Appendix

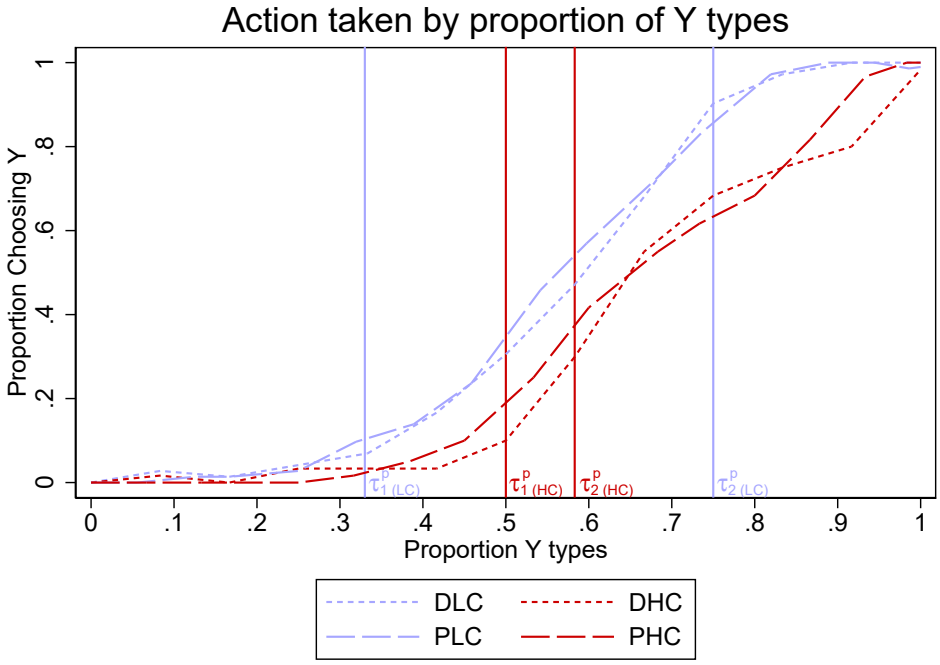

Figure 8: Average action taken by proportion of Y-types in the second part of each treatment, excluding non-transitioning groups. Vertical lines are the proportions at which transitions are predicted to begin and end.

Table 8: Group-level Choices (part 1 only)

	Prop. choosing $a_i = Y$	Prop. choosing $a_i = Y \mid \theta_i = Y$	Prop. choosing $a_i = Y \mid \theta_i = X$	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 9: Group-level Choices (part 2 only)

	Prop. choosing $a_i = Y$	Prop. choosing $a_i = Y \vert \theta_i = Y$	Prop. choosing $a_i = Y \vert \theta_i = X$	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 10: Group-level speed of transitions

	0% to 25%	0% to τ_1^p	25% to 100%	τ_1^p to τ_2^p	τ_1^p to 100%
DLC1	6	6	3	2	3
DLC2	5	5	6	4	6
DLC3	5	6	4	2	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	2.17	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	3	5
PLC2	8	8	4	3	4
PLC3	7	8	4	2	3
PLC4	9	10	3	1	2
PLC5	8	8	5	4	5
PLC6	6	7	6	2	5
PLC Mean	7.50	8.00	4.50	2.50	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 11: Speed of transitions, excluding non-transitioning groups

	0% to 25%	0% to τ_1^p	25% to 100%	τ_1^p to τ_2^p	τ_1^p to 100%
DLC	6.00	6.50	4.00	2.17	3.50
	(0.63)	(0.56)	(0.58)	(0.48)	(0.56)
DHC	8.60	9.00	1.20	0.60	0.80
	(0.93)	(1.14)	(0.20)	(0.25)	(0.20)
PLC	7.50	8.00	4.50	2.67	4.00
	(0.43)	(0.45)	(0.43)	(0.49)	(0.52)
PHC	9.60	10.60	2.40	0.20	1.40
	(1.25)	(1.25)	(0.25)	(0.20)	(0.25)
DLC - DHC	-2.60**	-2.50**	2.80***	1.57**	2.70***
	(1.09)	(1.20)	(0.66)	(0.57)	(0.65)
PLC - PHC	-2.10	-2.60**	2.10***	2.47***	2.60***
	(1.22)	(1.23)	(0.52)	(0.58)	(0.61)
DLC - PLC	-1.50**	-1.50**	-0.50	-0.50	-0.50
	(0.76)	(0.72)	(0.72)	(0.69)	(0.76)
DHC - PHC	-1.00	-1.60	-1.20***	0.40	-0.60**
	(1.56)	(1.69)	(0.32)	(0.32)	(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 12: Speed of transitions by number of Y-types

	0% to 25%	0% to τ_1^p	25% to 100%	τ_1^p to τ_2^p	τ_1^p to 100%
DLC	6.00	6.50	4.00	2.17	3.50
	(0.63)	(0.56)	(0.58)	(0.48)	(0.56)
DHC	9.33	9.50	0.83	0.50	0.67
	(1.05)	(1.06)	(0.31)	(0.22)	(0.21)
PLC	6.00	6.67	4.17	2.17	3.50
	(0.37)	(0.33)	(0.31)	(0.31)	(0.34)
PHC	8.67	9.33	1.67	0.00	1.00
	(1.33)	(1.17)	(0.42)	(0.00)	(0.26)
DLC - DHC	-3.33**	-3.00**	3.17***	1.67***	2.83***
	(1.23)	(1.20)	(0.65)	(0.53)	(0.60)
PLC - PHC	-2.67	-2.67	2.50***	2.17***	2.50***
	(1.38)	(1.22)	(0.45)	(0.31)	(0.43)
DLC - PLC	0.00	-0.17	-0.17	0.00	0.00
	(0.73)	(0.65)	(0.65)	(0.57)	(0.66)
DHC - PHC	0.67	0.17	-0.83	0.50	-0.33
	(1.70)	(1.58)	(0.52)	(0.22)	(0.33)

Values are the mean number 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. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$