THE ECONOMIC JOURNAL

The Economic Journal, **135** (*February*), 556–583 https://doi.org/10.1093/ej/ueae085
Advance Access Publication Date: 25 September 2024

STRATEGIC CONFORMITY OR ANTI-CONFORMITY TO AVOID PUNISHMENT AND ATTRACT REWARD*

Fabian Dvorak, Urs Fischbacher and Katrin Schmelz

We provide systematic insights on strategic conformist—as well as anti-conformist—behaviour in situations where people are evaluated, i.e., where an individual has to be selected for reward (e.g., promotion) or punishment (e.g., layoffs). To affect the probability of being selected, people may attempt to fit in or stand out in order to affect the chances of being noticed or liked by the evaluator. We investigate such strategic incentives for conformity or anti-conformity experimentally in three different domains: facts, taste and creativity. To distinguish conformity and anti-conformity from independence, we introduce a new experimental design that allows us to predict participants' independent choices based on transitivity. We find that the prospect of punishment increases conformity, while the prospect of reward reduces it. Anti-conformity emerges in the prospect of reward, but only under specific circumstances. Similarity-based selection (i.e., homophily) is much more important for the evaluators' decisions than salience. We also employ a theoretical approach to illustrate strategic key mechanisms of our experimental setting.

Conformity and anti-conformity are crucial processes contributing to social stability and at the same time promoting diversity on which societal dynamism depends. Various strands of the economic literature draw on the interplay of conformity and anti-conformity, for example, research on the exploration-exploitation dilemma in organisation theory (Schumpeter, 1934; March, 1991), on rational choice (Simon, 1955), on cultural-institutional evolution (Belloc and Bowles, 2013; Kets and Sandroni, 2021) or on cultural diversity (Kets and Sandroni, 2016). In these studies, conformity and anti-conformity are generally assumed to be an intrinsic tendency or preference that is independent of external factors.

People also conform to others' opinions in order to attract reward or avoid punishment (Festinger, 1953; Kelman, 1961; Allen, 1965; Amabile *et al.*, 1990; Bernheim, 1994; Shalley and Perry-Smith, 2001; Sakha and Grohmann, 2016). Our paper complements this literature by systematically investigating how evaluation affects *strategic* conformity as well as anti-conformity. Concretely, we show that the evaluation of individual behaviour by peers can provide incentives for conformist or anti-conformist *behaviour*.

* Corresponding author: Urs Fischbacher, Department of Economics, University of Konstanz, Universitätsstraße 10, 78464 Konstanz, Germany. Email: urs.fischbacher@uni-konstanz.de

This paper was received on 19 May 2022 and accepted on 18 September 2024. The Editor was Steffen Huck.

The data and codes for this paper are available on the Journal repository. They were checked for their ability to reproduce the results presented in the paper. The replication package for this paper is available at the following address: https://doi.org/10.5281/zenodo.13742789.

An earlier version of this paper circulated as *Incentives for Conformity and Anticonformity*.

Support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2117-422037984 is gratefully acknowledged. We thank Steffen Guth for excellent research assistance. We thank Samuel Bowles, Charles Efferson, Armin Falk, Sebastian Fehrler, Simon Gächter, Susanne Goldlücke, Ronald Hübner, Willemien Kets, Kerice Doten-Snitker, Simeon Schudy, Egon Tripodi, Ro'i Zultan, the research group of the Thurgau Institute of Economics (TWI) and the Department of Economics at the University of Konstanz, as well as participants of various conferences, workshops and seminars for helpful comments and suggestions. We acknowledge the feedback of anonymous referees and the Editor that has helped to improve the paper.

For the purpose of this study, we define conformity and anti-conformity as behavioural deviations from one's intrinsic choice preference due to information about others' choices. Building on Cialdini and Goldstein (2004), we refer to *conformity* as deviating from one's intrinsic preference *towards* the majority choice, and to *anti-conformity* as deviating from one's intrinsic preference *away* from the majority choice.

Our approach differs from and complements the research in the fields of biology and anthropology where conformity is considered as taking on the values, preferences and beliefs of others when these are common in a population (e.g., Boyd and Richerson, 1985; Bowles and Choi, 2013; Denton *et al.*, 2020). Instead, we focus on strategic conformist and anti-conformist behaviour, controlling for participants' intrinsic inclination to fit in or stand out.

Our notion of conformity as a behavioural response to social feedback relates to the economic literature on social learning and imitation (e.g., Vega-Redondo, 1997; Schlag, 1998; 1999; Apesteguia *et al.*, 2007) and information cascades (Banerjee, 1992; Bikhchandani *et al.*, 1992; Anderson and Holt, 1997; Guarino *et al.*, 2011). Different from these settings where decisions are typically backward looking, for example when people learn and imitate based on others' payoffs, we investigate settings in which people need to anticipate how others reward or punish, and in response, act strategically in conformist and anti-conformist ways.

People often evaluate others' actions and intentions to decide who deserves punishment or reward. This is particularly common at the workplace. For example, a team leader has to select one employee for an unpleasant job or, if the company plans layoffs, who will be fired. Typical cases where evaluations result in positive outcomes are job interviews, awards, promotions or competitions involving subjective selection criteria (e.g., in architecture or arts).

We study how people strategically respond to an evaluator's expected selection decision, considering two rules that the evaluators may apply. The first rule is salience, i.e., evaluators' attention may be drawn to the person standing out, determining whom they select. In anticipation, people may attempt to hide behind the majority under the threat of punishment (think about the allocation of an unpleasant task in a team meeting), but try to stand out when a reward is in prospect, as illustrated by Rubinstein (2013, p.195f):

> What do you recommend wearing to a job interview? No question, I am the right person to answer this question. I have never given a lecture with a jacket and a tie. I would argue that wearing jeans and a t-shirt is your dominant strategy: If you are a good student, then a department that will not give you a job because of your 'sloppy' appearance does not deserve to have you. If you are mediocre, then there are many other candidates like you and dressing casually is the only way for you to get noticed.

Salience has been established as an important mechanism in biology[1] and, more recently, also in economic decisions (as reviewed by Bordalo *et al.*, 2022). The experimental study on salience in reward and punishment contexts most closely related to our paper is Griskevicius *et al.* (2006). They found that a self-protection mindset induces conformity, while a mate-attraction mindset induces anti-conformity in men (but not in women).[2]

[1] Research in biology has shown that individuals' visual similarities or spatial proximities may lead to positive or negative consequences. According to the selfish-herd hypothesis, individuals reduce their risk of dying when forming groups as the risk that a specific individual of the collective is taken by a predator is distributed over all individuals (Hamilton, 1971; Vine, 1971). Among fish and crabs, the average nearest-neighbour distance drops sharply if individuals believe that an immediate threat is present (Viscido and Wethey, 2002; Sosna *et al.*, 2019) and increases if individuals are exposed to food cues (Schaerf *et al.*, 2017). In response to potential benefits, individuals across many species actively express their identity, relying on distinctive cues (Tibbetts and Dale, 2007). Standing out has been shown to be crucial for mate attraction, where differentiating oneself from rivals is key to success (Simpson *et al.*, 1999; Buss, 2003).

[2] Griskevicius *et al.* (2006) first asked their participants to rate how aesthetic they find a series of images. After having been primed towards either self-protection or mate-attraction, participants entered a computer chat with alleged

The second potential selection rule is homophily, i.e., the tendency of evaluators to more likely appreciate 'those who are alike in some designated respect' (Lazarsfeld and Merton, 1954, p.23).[3] Responding to homophily would imply wanting to appear similar to the evaluator to increase the chances of a reward and avoid punishment. The concept of homophily has widely penetrated the social sciences (see McPherson *et al.*, 2001 and Ertug *et al.,* 2022 for surveys), and there is evidence that homophily may benefit individuals who are targets of evaluations. For example, Mäkelä *et al.* (2010) documented a systematic similarity bias in managers' decisions when selecting employees who deserve to be promoted as 'talents'; Opper *et al.* (2015) showed that homophily increases recruitment chances to China's supreme decision-making body; and similarity between venture capitalists and founders or company executives positively influences funding decisions (Matusik *et al.*, 2008; Hegde and Tumlinson, 2014).

1. A New Methodological Approach

Though studying conformity experimentally has a long tradition across the social sciences, the methodological approach has been challenging. The pioneering literature in psychology has heavily relied on the debated use of deception (e.g., Asch, 1951; 1952; Crutchfield, 1955; Hertwig and Ortmann, 2001). An alternative, widely used method has been to present participants the same choice option twice, with and without information about others' decisions (e.g., Griskevicius *et al.*, 2006; Robin *et al.*, 2014; Amini *et al.*, 2017), bearing the confounds of a preference for consistency (Soll and Larrick, 2009; Falk and Zimmermann, 2017; Fehrler and Hughes, 2018) or an experimenter demand effect.

The literature on social influence has mainly focused on conformity, for the most part neglecting its vital complement of anti-conformity—though understanding their interplay matters for various domains, for example, the adoption of innovations (Griliches, 1957; Young, 2009) or in opinion dynamics.[4] Studies on anti-conformity are scarce (notable exceptions include Fromkin, 1970; Lynn and Harris, 1997; Ariely and Levav, 2000; Imhoff and Erb, 2009; Touboul, 2019) and particularly challenging—not only because anti-conformity is rare, but also because it is difficult to disentangle anti-conformity from independence (i.e., behaviour unaffected by social influence; Argyle, 1957; Crutchfield, 1962; Willis, 1963; Willis and Levine, 1976).

Our Experiment 1 introduces a novel experimental technique to identify conformist and anti-conformist behaviour based on transitivity. We first elicit individuals' preferences in two choices X versus Y and Y versus Z in the absence of information about others' choices. Employing transitivity, we then predict the choice of X versus Z, and we compare this prediction to

others they thought they would have a face-to-face discussion about aesthetic preferences with later. In the chat room, participants again had to rate one of the previous images, but this time publicly after being informed about the others' alleged ratings (which were pre-programmed to be either uniformly positive or negative).

[3] The term originates from the Ancient Greek *homós* (same, common, similar) and *philía* (love), and we use homophily in its literal sense (unlike a vast stream of literature on homophily that focuses on the formation of ties based on similarity, e.g., Currarini *et al.*, 2009; Golub and Jackson, 2012; Baccara and Yariv, 2013; Goldberg and Stein, 2018). Our notion of homophily also closely relates to the literature on similarity-attraction theory (starting with Byrne, 1971), which rests on the idea that people have positive feelings towards others who are similar. Far-reaching consequences of this powerful mechanism have been documented in the extensive literature on in-group favouritism and social identity, typically relying on the minimal group paradigm (as initiated by Tajfel, 1970 and Turner *et al.,* 1979; see Hewstone *et al.,* 2002 for a review), as well as in the economic literature on taste-based discrimination (Becker, 1971; Riach and Rich, 2002; Bertrand and Duflo, 2017).

[4] A single anti-conformist response can break unanimity, which is a particularly strong determinant of conformity in subsequent choices (Asch, 1955), and can prevent information cascades or influence the polarisation of opinions (Siedlecki *et al.*, 2016; Juul and Porter, 2019).

the individual's actual choice when being informed about others' choices among X versus Z. Participants make these choices in two domains, namely, answering questions about objective facts and expressing their subjective tastes over art paintings. Conformity is captured by adjustments of choices *towards* the majority choice, and anti-conformity by adjustments *away from* the majority choice. This technique, not only addresses the limitations of earlier methods, but it is crucial to cleanly separate conformist and anti-conformist choices from those that might appear socially influenced, but are actually independent (see Nail *et al.*, 2013 for a discussion).

Conformity has been documented to be a strong force in human interactions, while anti-conformity is much more rare—and this is also what we observe in Experiment 1. However, encouraging uniqueness is an important goal in the literature on organisations, and creativity has been linked to the degree of (non-)conformity in a society. For example, Shane (1992) showed that individualistic countries are more inventive than more conformist societies. Similarly, according to Goncalo and Staw (2006), individualistic rather than collectivist values foster creativity.

Experiment 2 enters the domain of *Creativity* and implements conditions inviting anti-conformity to be potentially expressed (more). Because creativity cannot be studied in a binary setting, we develop another novel, multi-dimensional setup featuring a colour creation task, where participants create several colours and choose which one to be published. This setup also serves to see how the mechanisms of the incentive structures in the binary choice setting of Experiment 1 generalise to a multi-dimensional choice environment.

We investigate the effect of evaluation on conformist and anti-conformist behaviour in these two laboratory experiments, comprising twenty treatments with a total of 871 participants. The general pattern of our implementation is as follows. Participants first make choices without knowing how others decide. Then, they are informed about their group members' decisions and make additional choices. A third party evaluates the group's choices by selecting one choice for reward (resulting in a payoff increase) or punishment (resulting in a payoff deduction), depending on the treatment. We also manipulate the relevance of salience versus homophily in three treatment variations.

We expect the possibility of punishment to induce strategic conformity, and the prospect of reward to limit conformity and induce strategic anti-conformity. The three domains of our experiments (objective facts, subjective tastes over art paintings, creativity) vary in the degree to which they may foster anti-conformist as opposed to conformist choices. We expect a shift away from conformity across those three domains. Treatments without evaluation serve as controls to elicit intrinsic preferences for conformity and anti-conformity in the absence of strategic incentives.

The central result of the experiments is that people show strategic conformity to avoid being punished, and they conform less when facing rewards. Strategic anti-conformity to attract a reward is rare and occurs only under certain conditions.[5] This is well in line with the evaluators' behaviour, creating incentives for conformity if the consequence for being selected is punishment, and under certain reward conditions creating incentives for anti-conformity. Evaluators' selection decisions are mainly driven by homophily. Salience plays a minor role and turns out to be relevant only in treatments where this mechanism is made very salient to participants.

[5] Our finding of ample conformist and rare anti-conformist behaviour is in line with the evolutionary literature showing that copying the behaviour of others is a superior strategy (Rendell *et al.*, 2010), while deviating from one's group can threaten group membership and lead to ostracism (Mahdi, 1986; Boehm, 1993; 2000; Wiessner, 2002; Boyd *et al.*, 2010).

Across our domains, conformity decreases and anti-conformity increases from objective facts over arts taste to creativity.[6] The mechanisms evoked by the incentive structures apply to our binary as well as multi-dimensional choice settings. Moreover, we observe individual heterogeneity in strategic conformity and anti-conformity.[7]

Next, we employ a theoretical framework to illustrate how punishment and reward affect conformity and anti-conformity under the salience and the homophily evaluation rules.

2. Theory

In this section, we sketch our model, which draws on the incentive structure of Experiment 1, capturing binary choices subject to evaluation. In Online Appendix A we provide the formal setup, including propositions and proofs.

There are three group members A_1, A_2 and B who choose between two options X and Y. Group members A_1 and A_2 decide simultaneously, whereas B takes an informed decision, being informed by the choices of A_1 and A_2. Then, an evaluator E observes the three decisions without knowing which decision is informed and selects one of the three group members. Depending on the treatment, the selected person receives a reward or a punishment m.

Each player prefers one of the two options. These preferences are positively correlated, and the correlation is related to a parameter $p \in \left(\frac{1}{2}, 1\right]$.[8] The incentives for A_1, A_2 and B consist of the reward or punishment and a utility $\tau_i, i \in \{1, 2, B\}$, if the player chooses according to the own taste.

The evaluator decides according to a rule, which can be either salience or homophily. The salience-based rule implies that he selects the minority player if there is one; otherwise, he chooses a player randomly. The homophily-based rule implies that, if possible, the evaluator rewards someone who has chosen in accordance with his own taste and punishes someone who has chosen against his taste (he randomises if this is not possible).[9]

The model predicts that if evaluators apply the salience rule, punishment incentivises conformist and reward incentivises anti-conformist choices. This is because conformity avoids being singled out, which is advantageous when the singled-out person is punished. Anti-conformity ensures to stand out, which is advantageous for reward.

If the evaluation is based on homophily, again, punishment incentivises conformity. Reward leads to less conformity than punishment and may evoke anti-conformity. The mechanisms are as follows: players have to assume that the evaluator's taste is likely to match the majority's choice. In the case of punishment, conforming to the majority is optimal if τ_B is not too high. The probability that the evaluator's taste corresponds to the minority is smaller than $\frac{1}{2}$. Thus, in the minority, the

[6] The observation of highest conformity when answering knowledge questions is consistent with the literature on social learning (e.g., Banerjee, 1992; Bikhchandani *et al.*, 1992; Lee, 1993; 1998; Anderson and Holt, 1997; Vives, 1997; Smith and Sorensen, 2000; Banerjee and Fudenberg, 2004). We also find conformity in the arts setting (even under reward), which is in line with the literature on frequency-dependent social learning (Boyd and Richerson, 1982; 1985; Efferson *et al.*, 2008; McElreath *et al.*, 2008).

[7] Our finding on heterogeneity in strategic conformist and anti-conformist behaviour is in line with studies on heterogeneity in preferences for conformity or anti-conformity (Argyle, 1957; Brehm, 1966; Jones, 1984; Corazzini and Greiner, 2007; Wright *et al.*, 2009; Jones and Linardi, 2014; Goeree and Yariv, 2015; Fatas *et al.*, 2018), which have been related to both individual traits (Fromkin, 1970; Lynn and Harris, 1997; Ariely and Levav, 2000; Imhoff and Erb, 2009) and cultural variation (Bond and Smith, 1996; Cialdini *et al.*, 1999; Kim and Markus, 1999; Yamagishi *et al.*, 2008).

[8] The probability that two players prefer the same option equals $p^2(1-p)^2$.

[9] We also discuss a rule based on performance in Online Appendix A for interested readers. This rule is secondary as it is not in the core interest of our study and applies only to specific environments in our setting.

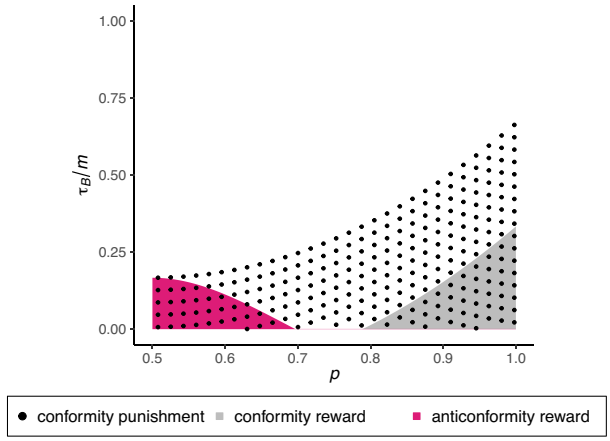

Fig. 1. *Limits for Conformity and Anti-Conformity in the Case of Homophily-Based Punishment and Reward.*
Note: The areas show combinations of p and τ_B/m for which there is conformity or anti-conformity in the respective treatments.

punishment probability is greater than $\frac{1}{2}$ and it is only $\frac{1}{3}$ when joining the majority. The case of homophily-based reward contains two countervailing effects. First, the evaluator is more likely to reward a majority choice, fostering blending in. Second, the probability of rewarding a majority choice is shared among all majority players—which provides a rationale for standing out. Thus, optimal behaviour depends on the correlation structure of preferences in the population.

Figure 1 shows for the homophily rule how the decision of player B depends on τ_B and p. Conformity and anti-conformity are not possible for high τ_B. In the punishment treatment, conformity is possible for any correlation parameter p, but it is more likely when the preferences are more strongly correlated. In the reward treatment, conformity is possible for strong correlations and anti-conformity is possible for weak correlations.

In a nutshell, if evaluation is based on salience, punishment incentivises conformity and reward incentivises anti-conformity. If evaluation is based on homophily, punishment incentivises conformity. Reward leads to less conformity than punishment and can even evoke anti-conformity. The model shows that evaluation has the potential to induce strategic conformity and anti-conformity in the sense that both can be rational responses to social influence in order to avoid punishment and attract reward.

3. Experiment 1

We investigate strategic conformity and anti-conformity in various settings that are determined by three dimensions: choice domains (i.e., *Facts* and *Taste*), incentives (i.e., *Reward*, *Punishment* and *Control* treatments) and the importance of salience-based evaluation. Table 1 provides an overview of our experimental setup, and the three dimensions are explained in the remainder of this section.

Our main interest is in strategic conformist and anti-conformist decisions, as captured by the *Reward* and *Punishment* treatments. They serve to measure whether and to what extent

Table 1. *Overview of the Setup of Experiment 1.*

Domains (within subjects)	Facts and Taste							
Incentives (between subjects)	Reward			Punishment			Control	
Salience (between subjects)	S0	S1	S2	S0	S1	S2	S0	S2
Pre-stage: salience training								
Payment for successful coordination on separate choice sets	–	–	✓	–	–	✓	–	✓
Stage 1: uninformed choice								
Participants go through twenty binary choice sets in the absence of information about others' choices	✓	✓	✓	✓	✓	✓	✓	✓
Stage 2: informed choice								
Participants go through ten binary choice sets, knowing how their group members have decided in this situation	✓	✓	✓	✓	✓	✓	✓	✓
Stage 3: evaluation								
The evaluator selects one of the group's choices	✓	✓	✓	✓	✓	✓	–	–
Group members' incentives								
Selected group member is rewarded	✓	✓	✓	–	–	–	–	–
Selected group member is punished	–	–	–	✓	✓	✓	–	–
Evaluator's (salience) incentives								
Evaluators are paid for successful coordination with other evaluators	–	✓	✓	–	✓	✓	–	–

positive and negative incentives affect behavioural adjustments towards or away from the majority, as detailed in Section 2. Control treatments without evaluation serve to elicit preferences for conformity and anti-conformity in the absence of strategic incentives. The purpose of the domains is to explore how incentives interact with the objective (*Facts*) and subjective (*Taste*) nature of the choice environment. Finally, given the low relevance of salience as compared to homophily determining evaluators' choices in our initial treatments (labelled *S0*), we add treatments *S1* and *S2* inducing more salience-based evaluations to conclusively understand whether or not this mechanism is relevant in such settings.

3.1. *Choice Domains: Facts and Taste*

Experiment 1 captures two choice domains, differing in the degree to which they may foster anti-conformity as opposed to conformity: *Facts* and *Taste*. In the *Facts* domain, participants face a series of difficult factual questions that have an objectively correct answer, though the data underlying the answers are very similar for both options and beyond the general knowledge of typical university students.[10] The objective, fact-based nature of this domain may allow for social learning, and the difficulty gives room for conformity to emerge.

In the *Taste* domain, participants choose among two similar art paintings.[11] As this domain involves subjective art preferences rather than objective facts, there is no right or wrong answer,

[10] Examples are as follows. 'Which country is older: Ghana or Niger?' Ghana was founded in 1957, one year before Niger was founded in 1958, and is therefore the correct answer. 'Who has sold more records in Germany: Britney Spears or Bon Jovi?' At the time of data collection, Bon Jovi was with 5,150,000 records sold, slightly ahead of Britney Spears with 5,050,000 records sold. 'Which airport had more passengers in 2014: Aeropuerto Madrid Barajas or Miami International Airport?' Madrid was somewhat more busy with 41,822,863 passengers, compared to Miami with 40,941,879 passengers.

[11] For example, two variants of the *Garden of the Artist* by Monet, two variants of a bride and a groom by Chagall, or flowers in a vase by van Gogh and Renoir. Full lists of the factual questions and paintings used in Experiment 1 are provided in Online Appendix C.

and we therefore expect less conformity compared to the *Facts* domain. Note that, based on our model as explained in Section 2, we nevertheless expect some degree of strategic anti-conformity in the *Facts* domain under *Reward* incentives.

3.2. *Eliciting Conformity and Anti-Conformity Based on Transitivity*

We measure conformity by the frequency of adjustments of choices *towards* the majority, and anti-conformity by the frequency of adjustments *away from* the majority. In groups of three, participants first make choices without knowing how others decide. Then, they are informed about others' decisions and choose again. We measure conformity and anti-conformity by comparing the choices without and with information about others' choices.

Adjustments occur if participants deviate from their intrinsic preference that we predict based on transitivity, using triplets of choice alternatives. Each triplet (X, Y, Z) generates three binary choice situations (X versus Y, Y versus Z and X versus Z). In Stage 1, each group member faces two out of these three possible binary choice situations, which are assigned such that the remaining binary choice situation is different for each group member. For example, if participant 1 (referred to as P1) faces the option pairs X versus Y and Y versus Z, P2 faces X versus Y and X versus Z, and P3 faces X versus Z and Y versus Z. Participants decide without being informed of the choices of their group members (referred to as uninformed choices).

In Stage 2, participants decide in the remaining binary choice situation they have not faced yet (X versus Z for P1, Y versus Z for P2, X versus Y for P3). Before deciding, they are informed about the Stage 1 decisions of their two group members in the same binary choice situation. In our example, before choosing between X and Z in Stage 2, P1 is informed about P2's and P3's decisions facing X versus Z. Thus, we refer to the Stage 2 decisions as informed choices.

To detect conformity and anti-conformity, we predict a participant's choice in Stage 2 assuming transitivity over the three items based on the two uninformed choices in Stage 1, and compare it to the actual informed choice in Stage 2. We also elicit their strength of preferences on a continuous scale (see Online Appendix C): after each uninformed choice, participants indicate how much they prefer their selected item over the alternative. This measure, not only serves to better understand intransitivity, but also to predict the uninformed choices in cases where in Stage 1, a participant prefers X over Y as well as X over Z.

For each of the *Facts* and *Taste* domains, we use ten sets of triplets per group. Accordingly, Stage 1 consists of twenty uninformed binary choice scenarios, and another ten informed binary choice scenarios in Stage 2. To provide an incentive for selecting according to their actual arts preferences, participants received an art postcard of one of their chosen paintings.[12] To provide some incentive for answering the factual questions correctly, at the end, participants were shown the correct solutions along with their answers.

3.2.1. *Discussion on transitivity: merits and limitations*

Measuring conformity (and, to a lesser extent, anti-conformity) experimentally has a long tradition across the social sciences, starting with Asch (1951). Typically, these phenomena have been studied by instructing confederates to give false answers to questions that do have an obviously correct answer, inducing a sharp mismatch between the subject's objective observations and the judgements of the pre-instructed group of confederates (as in Asch's experiments); or by using

[12] For each group within a session, a different set of postcards was randomly selected for being handed over, such that choosing a unique postcard would also apply to the entire session.

other forms of deception like mimicking others' responses by an apparatus (e.g., Crutchfield, 1955). For good reasons, the use of deception is meanwhile heavily debated, particularly in economics (Hertwig and Ortmann, 2001).

An alternative approach has been to present participants the same choice option twice, with and without information about others' decisions (e.g., Griskevicius *et al.*, 2006; Robin *et al.*, 2014; Amini *et al.*, 2017). Facing the same choice set twice implies that participants are likely to remember their uninformed first choice, which is often intended. This carries two potential confounds. First, people may have a preference for consistency (see Soll and Larrick, 2009; Falk and Zimmermann, 2017; Fehrler and Hughes, 2018) and thus stick to the same option in their informed second choice, which may reduce the observed effect of social influence. Second, asking the same question twice may trigger an experimenter demand effect that can go in either direction.

Our technique to measure conformity and anti-conformity using transitivity mitigates these concerns. In our setting, participants never face the same pairs of options twice. Even though our design does not fully exclude the possibility that participants may remember their uninformed choices, they would have to triangulate the informed choice that would be consistent with their two uninformed choices in order to make a deliberate consistent choice, or to respond to a perceived experimenter's demand. Given that participants go through a series of twenty uninformed choices, remembering each pair as a basis for triangulating is unlikely. If participants attempted to be consistent by remembering all choices they made in Stage 1 and triangulating their choices in Stage 2, the deviations we observe in the experiment would reflect lower bounds of conformity and anti-conformity.

The design of Experiment 1 accounts for violations of transitivity (Tversky, 1969; Loomes *et al.*, 1991). Our measures of conformity and anti-conformity—estimated in situations where a participant faces identical choices by the other group members in Stage 2 (referred to as majority information)—are corrected for baseline intransitivity—estimated from situations in Stage 2 where a participant faces two different choices of the other two group members (referred to as mixed information).

The transitivity approach does not work reliably when subjects (strategically) misreport their uninformed choice. According to our theoretical model, this can be the case if the *Reward* evaluation is based on salience. Since players should shift to their actually preferred choice in the mixed information scenario in Stage 2, we can assess this potential problem by comparing the intransitivity levels across the treatments and scenarios (see Section 5 below).

In essence, our new approach to measure conformity and anti-conformity based on transitivity reduces the potential confounds of consistency and an experimenter demand effect, and it is robust to violations of the transitivity assumption.

3.3. *Reward and Punishment Treatments*

The setup described above reflects our *Control* treatment (C), where participants' choices are disclosed to the other group members without any monetary incentives. To study strategic incentives for conformity and anti-conformity, we implement *Reward* and *Punishment* treatments, and we study how they interact with the *Facts* and *Taste* domains.

One of the three group members is assigned a bonus (*Reward*) or a deduction (*Punishment*) based on the following procedure. In Stage 3, each participant takes the role of an evaluator who picks one of the choices of another group. Evaluators are shown the three group members'

chosen options of a binary choice set (e.g., *X* versus *Y*). Thus, they either face three copies of the same option (e.g., *X*, *X*, *X*), or they see one option twice and the other option once (e.g., *X*, *X*, *Y*). By design, a set comprises two uninformed choices from Stage 1 and one informed choice from Stage 2, though this nature of a choice is not revealed to the evaluator. They select one of the three choices (shown in a randomised order) and, depending on the treatment, the corresponding group member receives a reward or punishment.

Each evaluator makes thirty such decisions in Stage 3. To investigate the relevance of homophily for the evaluators' decisions, the evaluation situations they face are also derived from triplets they have faced themselves when deciding in the role of a group member.

In the *Reward* treatments, 10 euros are added to the final payoff of the selected participant, while in the *Punishment* treatments, 10 euros are deducted. Participants receive a flat payment of 16 euros in the *Control*, 30 euros in the *Punishment* and 20 euros in the *Reward* treatments. The average payoff in the *Control* treatment is lower because these sessions were shorter as they did not include Stage 3, such that the treatments are comparable in their hourly payment. The flat payments are higher in the *Punishment* than in the *Reward* treatments to ensure equivalent lowest payoffs (20 euros in both cases) as well as similar average earnings. The fact that the *Reward* treatment incurs gains whereas the *Punishment* treatment incurs losses is an inherent feature of our design.[13]

For each domain, one evaluation decision per group was randomly selected for payment.[14] The *Reward* and *Punishment* treatments were implemented in a between-subject design, whereas the *Facts* and *Taste* domains were implemented within subjects. The order of the two domains was balanced across sessions. At the end of each domain, participants received feedback about the evaluation decisions in their group and their own payoffs.

3.4. *Salience Treatments*

As explained in Section 2, two potential mechanisms driving evaluators' decisions are homophily and salience. In the treatments presented so far, to our surprise, the data show little evidence for salience-based evaluations. We therefore implemented two more treatment variations pushing the salience rule.

In the *S0* treatments (as described above in Section 3.3), evaluators are not incentivised for their evaluation decisions. Applying coordination as a standard method to study salience (Mehta *et al.*, 1994a,b), in the *S1* and *S2* treatments, evaluators are incentivised to coordinate their evaluation decisions. Several participants evaluate the same choices (as in *S0*), but now, their payoff increases with each other evaluator who selects the same participant.[15] As there is no communication involved, coordination of the evaluation decisions has to be achieved by selecting the choice that is generally considered as salient.

[13] Our main focus is on how reward and punishment affect (anti-)conformity compared to an environment in the absence of such features, and we therefore implemented this more natural version of a control treatment (instead of having separate control treatments where reward and punishment would be allocated randomly at the end of a session).

[14] The main reason for paying only one instead of all choices is the experimental feature that participants receive a physical copy of their selected arts picture, and they should have the possibility to be unique in their entire session.

[15] An evaluator's payoff increases by 0.002 euros for each percentage point of the total number of other evaluators in the same session that match their decision.

Table 2. *Summary of the Treatment Data in Experiment 1.*

Domain	Facts								Taste							
Incentive	Reward			Punishment			No		Reward			Punishment			No	
Salience	S0	S1	S2	S0	S1	S2	S0	S2	S0	S1	S2	S0	S1	S2	S0	S2
Sessions	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Participants	54	51	45	54	51	39	60	42	54	51	45	54	51	39	60	42
Groups	18	17	15	18	17	13	20	14	18	17	15	18	17	13	20	14
Informed choices	540	510	450	540	510	390	600	420	540	510	450	540	510	390	600	420
Evaluations	1,620	1,530	1,350	1,620	1,530	1,170	1,800	1,260	1,620	1,530	1,350	1,620	1530	1,170	1,800	1,260

Note: The number of participants reflects the number of statistically independent observations in Experiment 1. All participants take the role of an evaluator and evaluations were elicited using the strategy method (Selten, 1967), which leads to the relatively large number of evaluation decisions.

In the *S2* treatments, we further induce salience-based evaluation by implementing a coordination training stage before the actual experiment starts. Participants are incentivised to coordinate their choices in a separate set of items, as explained in Online Appendix C.[16]

3.5. *Procedures*

We conducted a total of sixteen experimental sessions (six sessions of the *S0* treatments and four sessions of the *S1* treatments in 2017/2018, and another six sessions of the *S2* treatments in 2023) with 396 students at the University of Konstanz. Participants were recruited via ORSEE (Greiner, 2015) in the earlier sessions and via hroot (Bock *et al.*, 2014) in the later sessions. The experiments were conducted with z-Tree (Fischbacher, 2007), and instructions were shown as PDFs on participants' screens using E-nstructions (Schmelz, 2011). Participants' mean age was 22.3 years, and 54% were female. Table 2 summarises the numbers of sessions, participants, groups and choices in each treatment.

4. Experiment 2

Experiment 2 employs a substantially different and more complex setup than Experiment 1. We investigate responses to incentives for conformity and anti-conformity in the multi-dimensional setting of a creativity task, also departing from the binary nature of our theoretical model and from inferring participants' preferences based on transitivity.

To measure conformity and anti-conformity in a creativity task, we need to quantify how close or far two outcomes of a creative process are. To do so, we develop a new design making use of the RGB colour space. Participants design colours and choose one out of multiple self-designed options to be displayed to others, and evaluators reward or punish design choices.

We expect the colour creation task to induce more anti-conformity compared to the domains in Experiment 1 because designers exert an activity that is likely to be intrinsically motivating, they may identify with their creative output more than with a selection from a pre-defined choice set, and they have the possibility to express their identity and differentiate themselves from others by opting for a more unique colour.

Experiment 2 consists of four treatments following a 2 × 2 design, implemented between subjects, as shown in Table 3. Again, the main design dimension captures the incentives (*Reward* and *Punishment*), and we also increase the importance of salience-driven evaluation from

[16] Note that we also include an *S2* version of the *Control* treatment to see whether the coordination training has any effect on participants' choices in the first two stages. Obviously, we do not have an *S1 Control* treatment because there are no evaluators involved who could coordinate in Stage 3.

Table 3. *Overview of the Treatments in Experiment 2.*

Domain	Creativity			
Incentives (between subjects)	Reward		Punishment	
Salience (between subjects)	S0	S1	S0	S1
Pre-stage: colour generation				
All participants design colours for two minutes	✓	✓	✓	✓
Stage 1: uninformed choice				
Designers submit their pre-selected colour to their group of designers in the absence of information about others' choices	✓	✓	✓	✓
Stage 2: informed choice				
Designers can adjust their selected colour, knowing their group members' pre-selected colours	✓	✓	✓	✓
Stage 3: evaluation				
The evaluator selects one of the group's choices	✓	✓	✓	✓
Group members' incentives				
Selected group member is rewarded	✓	✓	–	–
Selected group member is punished	–	–	✓	✓
Evaluator's (salience) incentives				
Evaluators are paid for successful coordination with another evaluator	–	✓	–	✓

Note: These stages are repeated in each of the eight rounds of the experiment.

the *S0* treatments (eliciting baseline salience-based evaluations) to the *S1* treatments (adding coordination incentives for evaluators).

4.1. *Choice Domain: Creativity*

To let participants design colours, we developed a colour generation interface where each designer starts out with the same eight colours (red, green, blue, yellow, magenta, cyan, black and white), representing the vertices of the three-dimensional RGB colour space. Participants can then generate new colours by average mixing of two colours. Newly generated colours can be stored and reused to generate further colours. By repeatedly executing these steps, every colour in the RGB colour space can be approximated.[17]

At the beginning of each of the eight rounds, designers have two minutes to create new colours by mixing pre-existing colours. Their created colours from past rounds are available to them in future rounds. During this colour creation phase, designers can make a shortlist with up to four of their created colours.

4.2. *Eliciting Conformity and Anti-Conformity Based on Adjustments*

We elicit deviations from intrinsic preferences towards and away from the majority in Experiment 2 by comparing pre-selected options in the absence of information about others' choices with their adjustments following information about others' choices.

In groups of four, each designer first creates a private shortlist of their self-created colours (pre-stage colour generation), and then pre-selects one colour to be published within their group in Stage 1 (referred to as the uninformed choice). In Stage 2, after having seen the four pre-selected colours of the group, each designer has the opportunity to replace their pre-selected

[17] A picture of the colour generation interface is shown in Online Appendix C. For a movie illustrating the procedure of generating colours, see https://fdvorak.com/videos/creativity-task.mp4.

colour by a different colour from their shortlist (referred to as the informed choice). The final published colour set of a group consists of the uninformed Stage 1 choices of three designers and the informed Stage 2 choice of the fourth, randomly selected designer.

As designers generate their own colours, the choice alternatives differ across designers. To quantify the similarity or dissimilarity of a colour to the three pre-selected colours of the other group members, we use various measures based on the Euclidean distance (two different colour spaces and three ways to aggregate the distances to the other three colours; see Online Appendix B).

We consider a decision as conformist if the informed choice is closer to the colours of the other group members than the uninformed choice, and as anti-conformist if the informed choice is further away from the others' colours than the uninformed choice.

4.3. *Reward and Punishment Treatments*

In Stage 3, the three uninformed choices and the randomly selected informed choice are transmitted to an evaluator, who is naive with respect to this nature of a choice. The evaluator selects one of those four colours, and the corresponding designer receives a bonus (*Reward* treatment) or a deduction (*Punishment* treatment).

To implement reward and punishment, respectively, 2 euros are added and deducted, respectively, from the payoff of the selected designer. To ensure similar average earnings across treatments, designers receive a flat payment of 20 euros in the *Punishment* and 12 euros in the *Reward* treatments. Evaluators receive a flat payment of 16 euros. After each round, all designers are informed about the evaluation decision in their group, yielding the possibility to converge or diverge as a group over the course of the experiment.[18]

4.4. *Salience Treatments*

The experimental setup described so far reflects the *S0* treatments of Experiment 2. We also include *S1* treatments with coordination incentives, where two evaluators are assigned to the same group (instead of only one evaluator in *S0*), such that coordination is possible. Both evaluators receive an additional payment of 2 euros if their decisions coincide, and the decision of one randomly selected evaluator is implemented to determine the designer's payment of a given round.[19]

In Experiment 2, participants take the fixed roles of a designer or an evaluator. Evaluators are tied to a given group over the eight rounds to avoid spillovers from having evaluated other groups, which might bias their salience perceptions given the multi-dimensional nature of options. The same concern would apply had evaluators participated as designers themselves.

[18] We did not include a *Control* treatment without incentives in Experiment 2, as it is unclear what an appropriate *Control* treatment would be. Eliminating Stage 3 and going through Stages 1 and 2 without the outcomes being shown to an evaluator may feel odd and confusing to participants as there would be no purpose in doing so. Showing the outcomes to an evaluator in Stage 3 without implementing monetary incentives would lean towards the *Reward* treatment as the evaluator pays attention to the selected colours.

[19] When conducting the *S0* treatments, we did not anticipate observing so little salience-based evaluation, or that we would add treatments pushing this mechanism. So, it seemed natural to assign one evaluator to each group. Even though the designers' decisions are shown to one evaluator in *S0*, but to two evaluators in *S1*, designers' monetary incentives remain unchanged across these treatment variations.

Table 4. *Summary of Treatment Data of Experiment 2.*

Domain	Creativity			
Incentive	Reward		Punishment	
Salience	S0	S1	S0	S1
Sessions	4	4	4	4
Participants	115	120	120	120
Groups	23	20	24	20
Informed choices	736	640	768	640
Evaluations	184	320	192	320

Note: The number of groups reflects the number of statistically independent observations.

4.5. *Procedures*

Each session consisted of three parts. First, participants made their decisions in the experimental treatments as detailed above. Second, we elicited which colour is generally considered to be salient by performing a Krupka–Weber coordination task (Krupka and Weber, 2013) across the four colours shown to evaluators. Third, participants rated how beautiful and interesting they found each of those four colours on continuous scales ranging from zero (not beautiful/not interesting at all) to one (very beautiful/very interesting).

The experiment was conducted in sixteen sessions with 475 students at the University of Konstanz in 2018. Participants were recruited via ORSEE (Greiner, 2015), excluding participants who had participated in Experiment 1. The experiment was implemented in z-Tree (Fischbacher, 2007), and instructions were shown as PDFs on participants' screens using E-nstructions (Schmelz, 2011). The mean age was 21.3 years, and 63% were female. Table 4 summarises the numbers of sessions, participants, groups and choices in each treatment.

5. Experimental Results

We start out by showing our main results on responses to incentives for conformity and anti-conformity, and how these responses interact with our experimental settings. Then we turn to our findings on the importance of the homophily and salience rules driving the evaluators' decisions, which determine the actual incentives for conformity and anti-conformity in our data. Throughout this section, we present the results of our two experiments jointly.

5.1. *Responses to Incentives for Conformity and Anti-Conformity*

To depict the responses to incentives in our treatments, we rely on the conceptual framework provided by the Willis–Nail model of social response (Willis, 1965; Willis and Levine, 1976; Nail, 1986; Nail and Van Leeuwen, 1993; Nyczka and Sznajd-Weron, 2013; Nyczka *et al.*, 2018). The three vertices of the model space represent the three canonical responses to social influence: conformity (C), anti-conformity (A) and independence (I).

5.1.1. *Average responses to incentives*

Figure 2 shows average responses to incentives in our treatments according to this framework. We operationalise the horizontal independence dimension as the relative frequency of adjustments of the informed choices in either direction, i.e., the sum of the relative frequencies of adjustments

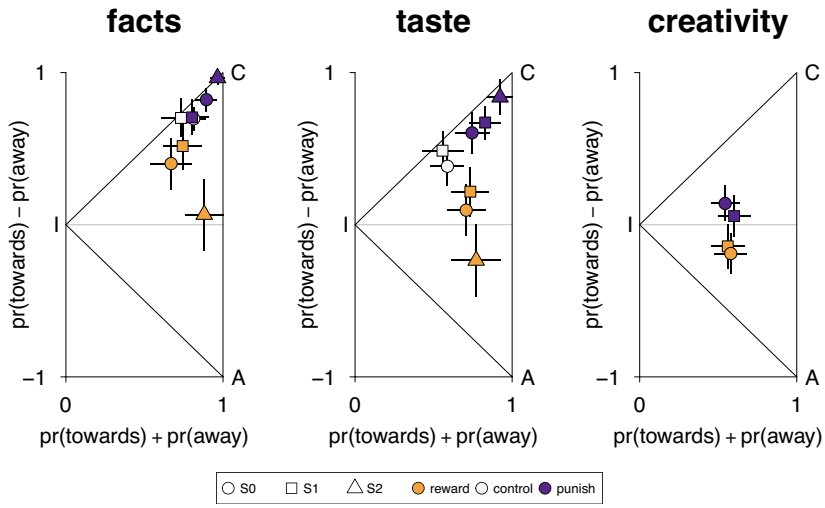

Fig. 2. *Average Response to Social Influence across Treatments.*
Note: The circle, square and triangle markers show average behaviour, and whiskers indicate 95% confidence intervals along the two model dimensions based on the *t*-distribution and block-bootstrapped SEs, where blocks are subjects in Experiment 1 and matching groups in Experiment 2. Results of the *Creativity* domain rely on the minimal Euclidean distance in the RGB colour space in the second half of the experiment. (The reason is that, by then, participants gained experience with the interface and created a desirable variety of colours, allowing them to respond to others' choices. Online Appendix Figure B2 shows that the treatment effects are robust when we use data from all periods as well as alternative measures of similarity.)

towards and away from the majority. The vertical conformity–anti-conformity dimension (vertices C and A) captures the net direction of the adjustments, i.e., the relative frequency of adjustments of the informed choice towards minus away from the majority.

Orange markers (lighter grey) capture the *Reward* treatments, purple markers (darker grey) the *Punishment* treatments, and white markers capture the *Control* treatments. Circles refer to our main *S0* treatments without coordination incentives. Squares refer to *S1* treatments, and triangles to *S2* treatments, fostering the salience rule in evaluations.[20]

Figure 2 conveys four results. First, and unsurprisingly, we observe ample conformity, with most of the markers in the upper halves of the triangles. Anti-conformity is rare, but exists in certain environments (indicated by the markers in the lower halves of the triangles). Second, comparing the points in orange (light grey) and purple (dark grey) reveals that experimental behaviour is consistent with our theoretical predictions: the prospect of *Punishment* creates incentives for conformity, whereas the prospect of *Reward* reduces incentives for conformity.

Third, as intended by our design, conformity tends to be stronger in the *Facts* domain than in the *Taste* domain, and is weakest in the *Creativity* domain, as shown by the markers moving away from the conformity vertex of the triangle from the left over the middle to the right panel.

[20] Throughout our figures, we report results using bootstrapped SEs, because we need to control for the potential statistical dependence of the choices made by the same participant in Experiment 1, and the choices made in the same matching group in Experiment 2. We report confidence intervals based on block-bootstrapped SEs because bootstrapping is straightforward to apply to more complex estimators such as those derived from mixture models. An alternative, statistically equivalent approach are cluster-robust SEs, which yield very similar results and do not affect our interpretations.

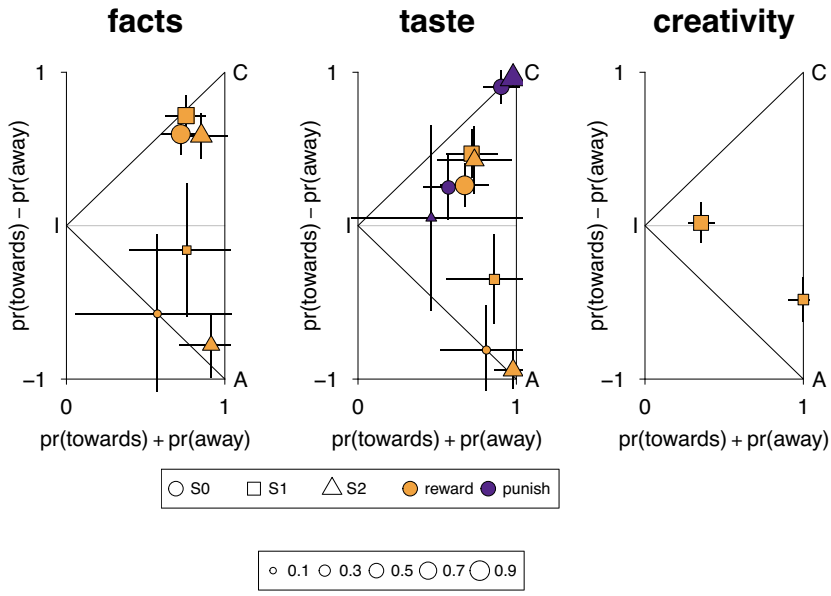

Fig. 3. *Heterogeneity in Types of Responses to Social Influence.*
Note: The circle, square and triangle markers indicate the average behaviour of the types, and their sizes capture the estimated frequency of the type in the sample. Whiskers indicate the 95% confidence intervals for each type based on the *t*-distribution and block-bootstrapped SEs. We choose the number of behavioural types for each treatment that minimises the Bayesian information criterion (Schwarz, 1978).

Fourth, similar behaviour in the *S0* and *S1* treatments suggests that coordination incentives to promote salience have little effect. Participants only respond to this rule, potentially driving the evaluators' decisions when it is made highly salient in the *S2* treatments, and the behavioural changes are in the expected directions.

5.1.2. *Heterogeneity in the response to social influence*

Figure 3 reveals that in some (but not all) settings, the averages in Figure 2 arise from actually very distinct types. Each marker indicates the response of a behavioural type in a given treatment. The number of behavioural types, their positions and their frequencies are estimated based on mixture models, using the R package *stratEst* (Dvorak, 2023). We use the Bayesian information criterion (Schwarz, 1978) to select the number of types. The estimated frequency of each type is represented by the size of its marker. If no markers are shown for a treatment, there are no individual differences, and the behaviour is adequately summarised by the average shown in Figure 2.

For the six *Reward* treatments of Experiment 1 (*Rewards S0*, *S1*, *S2* in the *Facts* and *Taste* domains), there are two behavioural types in each treatment that differ in their positions on the conformity–anti-conformity dimension. Most notably, we find an anti-conformist type in each *Reward* treatment, with estimated frequencies ranging from 15% to up to 51%. As expected, anti-conformity is most frequent and strongest in the *S2* treatments. Not surprisingly, conformist types are frequent with estimated shares ranging from 49% to 84%, and conformity is stronger in the *Facts* domain than in the *Taste* domain.

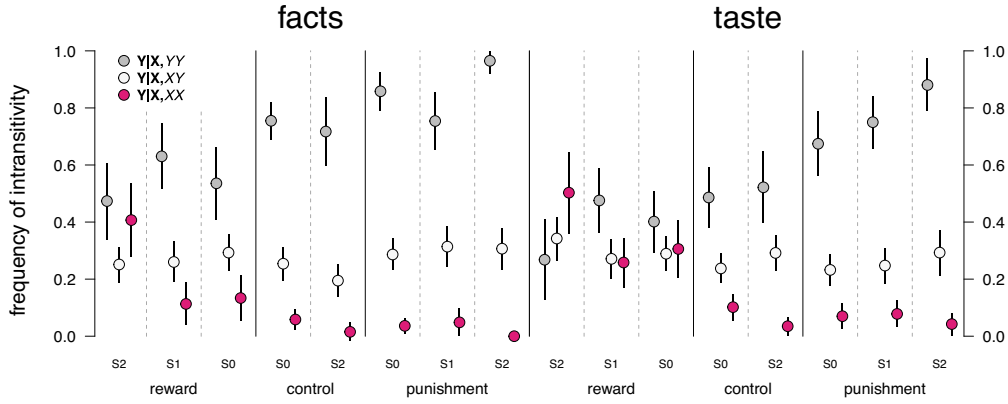

Fig. 4. *Intransitivity in the Informed Choices in Experiment 1.*
Note: Circles show the relative frequencies of intransitivity, conditional on the number of other group members with a choice different from the predicted choice. Whiskers indicate 95% confidence intervals of the estimates, based on block-bootstrapped SEs (10,000 samples, matching subject ID). Note that within each domain, the treatments are ordered according to our hypotheses with respect to the degree of conformity (increasing) and anti-conformity (decreasing).

In the *Creativity* domain, heterogeneity exists only in the *S1 Reward* treatment, with the most frequent type being independent (66%) and a minority type exhibiting substantial anti-conformity. In line with our expectations, the frequency of conformist responses to *Reward* decreases across our domains from *Facts* over *Taste* to *Creativity*.

Under *Punishment*, we observe heterogeneity only in the *Taste* domain in the *S0* and *S2* treatments: we find a highly conformist majority type (57% to 85%) and a less conformist minority type (15% to 43%), while anti-conformity is absent.

5.1.3. *Intransitivity in the informed choices of Experiment 1*

Figure 4 shows the relative frequency of intransitivity, i.e., deviations from the predicted choice when being informed about others' choices in Experiment 1—for example, selecting **Y** instead of the predicted choice **X**. There are three scenarios with respect to the number of other group members who chose in line with a participant's predicted choice.

The scenario where one other group member chose in line and one chose against a participant's predicted choice **X** provides the baseline, as shown by the open circles (**Y**|**X**, *XY*). The share of baseline intransitivity amounts to about 27% on average (ranging from 20% to 34%) and does not differ systematically across treatments. Accordingly, the concern that people may strategically misreport their actually preferred option in the uninformed decision under *Reward* does not seem to be very relevant.

Participants could adjust their informed choice towards the majority if both other group members selected *Y* (lighter grey filled circles, **Y**|**X**, *YY*), and away from the majority if no other group member selected *Y* (darker, red filled circles, **Y**|**X**, *XX*). To assess how much conformity and anti-conformity a treatment evokes, the baseline in Figure 4 is essential. A treatment induces conformity if the lighter grey filled circles exceed the frequency of baseline intransitivity, and anti-conformity is induced only if the darker, red filled circles *exceed* baseline intransitivity. Thus, our control treatments show a high level of conformity in the two domains—in terms

of increased conformity where conformity is possible (lighter grey filled circles are above the white circles), and in terms of reduced anti-conformity where anti-conformity would have been possible (darker, red filled circles are below the white circles).

In both domains, the prospect of *Punishment* mainly induces conformity, especially in the *S2* treatments, while it has little bearing on anti-conformity. In contrast, the prospect of a *Reward* reduces conformity, and increases anti-conformity in the *S2* treatment of the *Taste* domain. Thus, our data suggest that while the prevalence of conformity is always affected by the *Reward* and *Punishment* incentives, anti-conformity requires specific settings to occur.

5.1.4. *Similarity in the informed choices of Experiment 2*

In the *Creativity* domain, we investigate conformity and anti-conformity by measuring the degree of similarity between a designer's informed choice (in relation to the four possible options at hand) and the submitted colours of the other three group members. Each designer's shortlist of four colours yields four potential degrees of similarity to the others' colours.

To measure the degree of similarity, we calculate the Euclidean distance of a designer's colour to the colours of the other three group members. We find a higher degree of similarity of the adjusted colours to the colours of the other three group members in the *Punishment* compared to the *Reward* treatments. Online Appendix Table B5 shows that the average distance of the adjusted colours in the *Punishment* treatments is consistently smaller for all our distance measures (see Online Appendix B.1 for details on the six different distance measures we use).[21]

We shed light on the mechanisms behind these treatment differences by identifying the strategies used when adjustments occur. To do so, we calculate the rank of similarity for each colour in a designer's shortlist to the submitted colours of the other group members. The highest rank refers to the colour with the largest minimum Euclidean distance to the others' colours (i.e., lowest-rank colours are most similar and highest-rank colours are least similar).

Figure 5 reveals that the treatment differences are driven by different adjustment strategies. If adjustments occur, in the *Reward* treatments, designers most frequently adjust to the colour with the higher distance ranks 3 and 4 (adjusting away from the others), whereas in the *Punishment* treatments, participants most frequently adjust to the colours of the lower ranks 1 and 2 (adjusting towards the others). The salience-inducing coordination treatments *S1* do not affect these adjustments in a consistent way compared to the main *S0* treatments. In the *S0* treatments, the average ranks are 2.73 under *Reward S0* and 2.41 under *Punishment* (bootstrapped p-value $<$ 0.001). In the *S1* treatments, the average ranks are 2.61 under *Reward* and 2.43 under *Punishment* (bootstrapped p-value $= 0.026$).

5.1.5. *Determinants of the adjustment of informed choices*

What explains whether or not participants deviate from their predicted choice? We identify two determinants affecting the probability of intransitivity in the informed choice of Experiment 1 (as detailed in Online Appendix B.5). First, the stronger a group member's predicted preference strength in the predicted informed choice, the less likely intransitivity occurs in most of the

[21] Online Appendix Figure B3 also shows that the average distance decreases as participants gain experience over the course of the experiment in the *Punishment* treatments, reflecting increasing conformity. The distance remains rather stable in the *Reward* treatments (except for a drop in the last round).

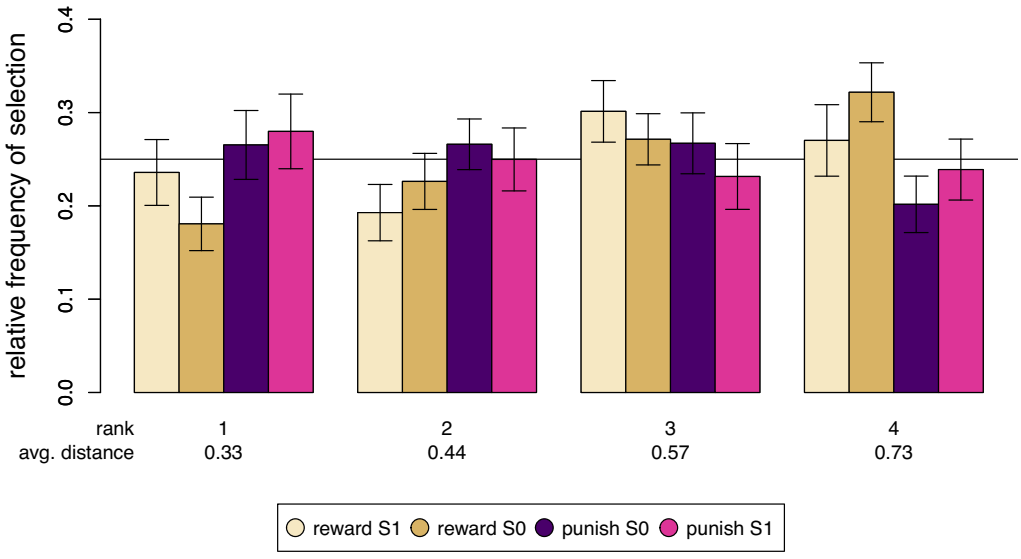

Fig. 5. *Frequency of Adjustments in Distance Ranks in Experiment 2.*
Note: The figure is based on adjusted choices only. The higher the ranks, the larger the minimum Euclidean distance of the selected colour to the others' colours. Whiskers indicate 95% confidence intervals, based on the *t*-distribution and block-bootstrapped SEs (10,000 samples, matching group ID).

treatments.[22] An exception is the *Facts Punishment S2* treatment where the majority choice is always selected (as evident from Figure 2), irrespective of the preference strength.

The second determinant refers to how common the predicted item is in the group. Overall, intransitivity is less likely if the predicted item matches the choices of the other group members, but more likely if the predicted choice stands out. The prevalence of conformity varies systematically across treatments in line with our theoretical predictions. Deviating from the predicted majority choice is much *less* likely in the prospect of *Punishment* than in the prospect of *Reward*, whereas deviating from the predicted unique choice is much *more* likely in the *Punishment* than in the *Reward* treatments. These patterns are particularly pronounced in the *S2* treatments.

In the *Creativity* domain, the decision to adjust the choice under social influence is predicted by strategic considerations in the *Reward*, but not in the *Punishment* treatments. Participants more frequently adjust their informed choice under *Reward* the more similar their initially chosen colour is to the colour of a group member. Designers' ratings of how beautiful and interesting they perceive their created colours do not predict the adjustment decision. Thus, in the *Creativity* domain, designers try to stand out if their choice is too similar to other choices in the prospect of a *Reward*. In the prospect of *Punishment*, whether or not adjustments are made is neither predicted by the similarity between the initially chosen colour to the others' colours, nor by the ratings of how beautiful and interesting the designer finds the initially chosen colour.[23]

[22] Preference strength is measured by the average of the signed preference strengths of the two uninformed choices. If σ_{XY} is the strength of the preference in favour of X when the alternative is Y then we predict σ_{YZ} as the average of σ_{YX} and $\sigma_{XZ} = -\sigma_{ZX}$.

[23] An explanation for the lack of explanatory power of the similarity between the initially chosen colour and the colour of a group member under *Punishment* could be that, depending on the choice alternatives, an adjustment to a similar

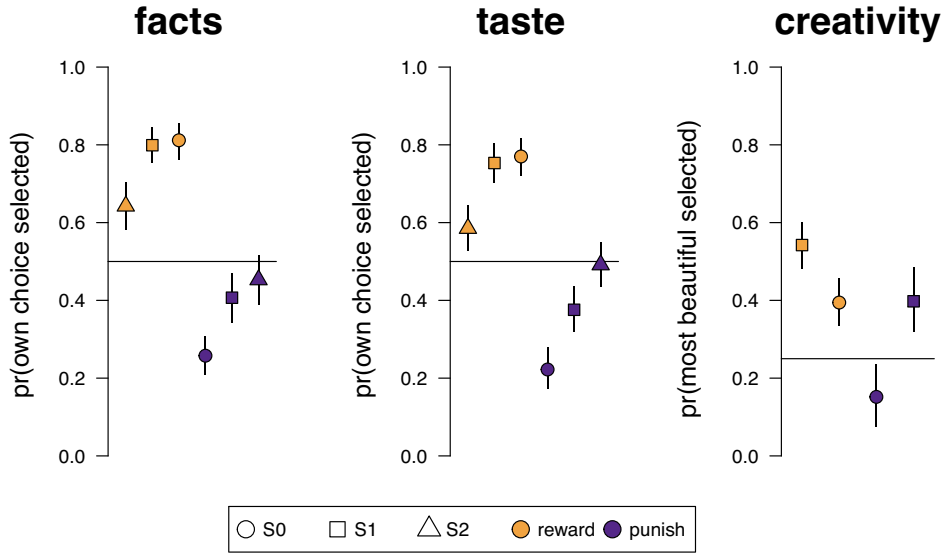

Fig. 6. *Selection Probabilities of the Item Matching the Evaluator's Own Preference.*
Note: In the left and central panels, a circle, square or triangle marker indicates the probability of the evaluator selecting the answer or question that matches their own prior choice in the role of a group member. In the panel on the right, a circle or square marker indicates an evaluator's probability of selecting the colour they rated as most beautiful. Whiskers indicate block-bootstrapped 95% confidence intervals based on the *t*-distribution and block-bootstrapped SEs. Horizontal lines mark the expected probability in the case of random selections.

5.2. *Evaluators' Decisions: Homophily and Salience*

As outlined in our model, homophily and salience may both drive the evaluators' decisions. Disentangling their relative importance is essential to understand the incentives for conformity and anti-conformity in our treatments (in particular in the reward treatments where predictions differ). Figures 6 and 7 give an impression of the importance of the two mechanisms.[24] We complement the figures with regression analyses.

5.2.1. *Evaluators' frequencies of rewarding and punishing based on homophily*

Figure 6 shows evaluators' selection decisions based on homophily across the treatments including evaluation. The left and central panels show the relative frequencies of selecting the answer and painting that matches the evaluators' choices when they themselves decided in the role of a group member. The panel on the right shows the relative frequencies of evaluators selecting the colour they rated as being most beautiful.

Homophily is very powerful in driving evaluators' selections. In the *Facts* and *Taste* domains, evaluators select the item matching their own prior choice for a *Reward* in about 80% of the cases

colour is not always possible. Because of the multi-dimensional choice space, it will usually be possible to adjust to a less similar colour under *Reward*, but not necessarily to a more similar colour under *Punishment*.

[24] Note that the percentages of the two figures may add up to more than 1 because the item matching the evaluator's own choice (homophily) may coincide with the single item (salience).

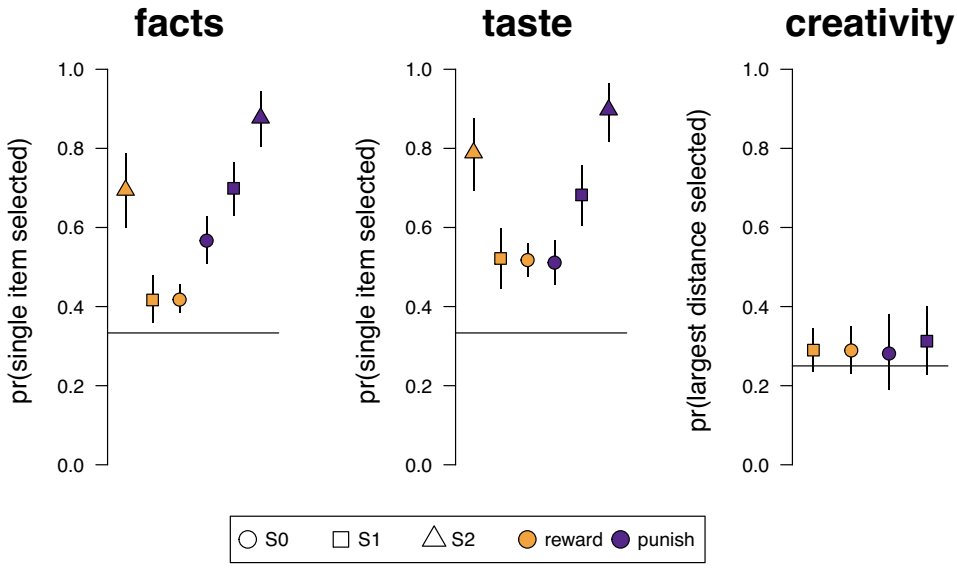

Fig. 7. *Selection Probabilities of the Item Standing Out.*
Note: In the left and central panels, a circle, square or triangle marker indicates the probability of the evaluator selecting the single answer or question. In the panel on the right, a circle or square marker indicates the evaluator's probability of selecting the colour with the largest minimal distance. Whiskers indicate block-bootstrapped 95% confidence intervals based on the *t*-distribution and block-bootstrapped SEs. Horizontal lines mark the expected probability in the case of random selections.

in *S0* and *S1*, and yet in about 60% in the *S2* treatments. Homophily is also the predominant mechanism for assigning a *Reward* in the *Creativity* domain.

When assigning *Punishment* in the absence of coordination incentives (*S0*), evaluators avoid the item they like themselves in all domains. This mechanism is mitigated in the *S1* and *S2* treatments.

5.2.2. *Evaluators' frequencies of rewarding and punishing based on salience*

Figure 7 shows evaluators' selection decisions based on salience. The left and central panels show the relative frequencies of selecting the answer and painting that is chosen by only one group member. The panel on the right shows the relative frequencies of evaluators selecting the colour with the largest minimal distance to the other group members' colours.

In the *Facts* and *Taste* domains, the frequency of the single item being selected is always larger than the one-third frequency in the case of random allocations, implying that standing out always increases the chance to attract a reward, whereas punishment can be avoided by 'hiding' in the majority.

In the absence of coordination incentives (*S0*), in the *Facts* domain, evaluators select the single answer more frequently when punishing instead of rewarding, whereas in the *Taste* domain, positive or negative consequences do not affect the probability of the single painting being selected. Moreover, while evaluators choose the single painting in slightly more than half of the cases in both *S0* treatments of the *Taste* domain, the single answer to *Facts* questions is selected more frequently under *Punishment* than under *Reward*. These observations are in line with

Table 5. *Salience-Based versus Homophily-Based Evaluation.*

	Facts						Taste						Creativity			
	Reward			Punishment			Reward			Punishment			Reward		Punishment	
	S0	S1	S2	S0	S1	S2	S0	S1	S2	S0	S1	S2	S0	S1	S0	S1
Salience	−0.50	−0.30	0.90	0.26	0.93	1.93	0.10	0.07	1.39	0.14	0.76	2.32				
	(0.09)	(0.10)	(0.09)	(0.09)	(0.08)	(0.09)	(0.09)	(0.09)	(0.08)	(0.09)	(0.08)	(0.09)				
Homophily	1.53	1.40	0.74	−1.07	−0.44	−0.28	1.21	1.12	0.56	−1.26	−0.57	−0.25				
	(0.13)	(0.13)	(0.11)	(0.12)	(0.12)	(0.11)	(0.12)	(0.11)	(0.11)	(0.13)	(0.11)	(0.10)				
Salient colour													1.11	1.25	0.25	1.17
													(0.18)	(0.15)	(0.23)	(0.17)
Beautiful colour													1.09	1.91	−2.15	0.35
													(0.55)	(0.41)	(0.58)	(0.40)
Interesting colour													0.56	1.07	−1.22	0.25
													(0.57)	(0.46)	(0.92)	(0.48)
Obs.	648	612	540	648	612	468	648	612	540	648	612	468	736	1,280	768	1,280
N	54	51	45	54	51	39	54	51	45	54	51	39	23	40	24	40
LL	−302	−303	−308	−366	−357	−177	−349	−342	−269	−342	−367	−142	−213	−308	−234	−383

Note: Conditional logit regression with the evaluator's selected item as the dependent variable. Coefficients and block-bootstrapped SEs are reported in parentheses. For the *Facts* and *Taste* domains, *salience* is a dummy indicating the single item, and *homophily* is a dummy indicating the evaluator's own prior choice. For the *Creativity* domain, the *salient colour* is identified by the minimum of the Euclidean distances to the other three colours. The variables reflecting how *beautiful* and *interesting* an evaluator perceives a colour are continuous. The labels *Obs.*, *N* and *LL* refer to the number of observations, the number of participants and the log likelihood of the model, respectively.

the interpretation that evaluators may perceive the less frequently chosen answer in a group as a negative signal about its correctness, while there is no truth involved in subjective arts preferences.

Comparing the *S0* and *S1* treatments reveals that coordination incentives increase the probability of the single item being selected only in the *Punishment*, but not in the *Reward* treatments of both domains. We return to this observation at the end of this section. Eventually, the *S2* treatments succeed in triggering the salience rule as evaluators assign *Punishment* to the single item in about 90% of their choices, and they assign *Reward* to the single answer in about 70% and to the single painting in about 80% of their choices.

In the *Creativity* domain (right panel of Figure 7), incentives for standing out are small as the relative frequencies of the evaluator selecting the colour with the largest minimal distance are not much larger than the one-fourth expected frequency in the case of random selections, and they are unaffected by the treatments. In what follows, we contrast the potential impact of salience with homophily on evaluators' decisions.

5.2.3. *Salience-based versus homophily-based evaluation*

To analyse the extent to which evaluators allocate *Reward* and *Punishment* based on salience or homophily, we rely on conditional logit models (McFadden, 1974) and regress the evaluator's selected item on characteristics of the evaluated choices. We use the R package *mlogit* (Croissant, 2019) to obtain maximum likelihood estimates for the coefficients and block-bootstrapped SEs. The regression results are shown in Table 5.

The models referring to Experiment 1 contain the dummy variables *salience* (indicating the single answer or painting) and *homophily* (indicating whether or not the selected item equals the evaluator's own prior choice in the role of a group member).

Concerning Experiment 2, *salient colour* is a dummy indicating the colour generally considered to be salient, as derived from a Krupka–Weber coordination task (Krupka and Weber, 2013) among all participants across the four colours. The variables capturing how *beautiful* and *interesting* a colour is refer to the evaluator's rating.

Table 5 confirms that homophily is a powerful and robust evaluation rule. The coefficients related to homophily (*Homophily* and *Beautiful colour*) are always positive for the *Reward*

treatments and largely negative for the *Punishment* treatments. The effect sizes are substantial and precisely estimated, especially across all treatments in Experiment 1, and stronger compared to the coefficients reflecting salience (particularly so in *S0*). An exception is the *Punishment S1* treatment in the *Creativity* domain, where homophily does not escape being selected for punishment. Overall, evaluators favour those who appear similar to themselves—they are more likely to receive a reward and less likely to be punished.

In contrast, salience-based allocation of reward and punishment is a much weaker mechanism. If punishment and reward were assigned based on salience, the coefficients of the corresponding variables should be positive. In our main *S0* treatments where coordination incentives are absent, the coefficients of the salience variables in Table 5 are generally small, indicating that salience plays a minor role. The coefficients for *salience* are even negative in the *Facts* domain, which is consistent with the interpretation that evaluators may attempt to punish wrong answers (believing that the majority answer is more likely to be correct).

In the *S1* treatments, the coefficients of the variables reflecting salience show that our experimental manipulation to induce salience-based evaluation worked in the *Punishment* treatments of Experiment 1 and Experiment 2, but failed in the *Reward* treatments of both experiments, corroborating our observation from Figure 7.

Two potential explanations may account for these different effects of coordination incentives in the *Punishment* and *Reward* treatments. First, to allocate *Punishment*, selecting the salient item and not the item matching the evaluator's own prior choice is plausible. However, evaluators may still prefer to reward someone who decides like themselves (i.e., based on homophily)—regardless of the coordination incentives meant to foster salience.

Second, while the *S1* treatments create incentives for evaluators to coordinate, salience in terms of the single item may not be the only coordination criterion. For example, in the *Punishment* treatment of the *Facts* domain, evaluators could also attempt to coordinate on the assumingly wrong answer. Even though we constructed the answer options such that the correct answers are hardly known by anyone, evaluators may perceive the frequency of a selected answer as a signal about its correctness. Thus, coordinating on the single item coincides with coordinating on the probably wrong answer and thus increases the probability of the single item being selected from *S0* to *S1*. In the *Reward* treatment of the *Facts* domain, the evaluation criterion is ambiguous: evaluators may coordinate on the single (but maybe wrong) answer, or on the probably correct (but majority) answer, which may explain that the *S1* treatment does not differ from *S0* in this case.

Adding the mechanistic element of training participants to coordinate on the single item in *S2* finally gets salience to work and diminishes the relative importance of homophily under *Punishment* as well as *Reward*, and in both domains of Experiment 1.

In a nutshell, evaluators' decisions suggest that homophily is a very powerful driving force, whereas salience needs to be very salient in order to take effect. Accordingly (and in line with the group members' choices described in Section 5.1), matching the evaluator's taste may be more important than standing out to avoid punishment, and in particular to attract a reward.

6. Discussion

The effect of incentives on the balance between conformist and anti-conformist behaviour has received little attention in the literature on social influence. We show theoretically and experimentally that evaluation can incentivise both conformist and anti-conformist behaviour. In a theoretical model we analyse how punishment and reward affect conformity and anti-conformity. For

both decision rules of the evaluator that we investigate—salience and homophily—punishment creates incentives for conformity. Reward compromises conformity and can create incentives for anti-conformity. In two laboratory experiments, we find that evaluation induces strategic conformity in the case of punishment, and may induce strategic anti-conformity in prospect of reward. The effects of evaluation are consistent across three choice domains, despite varying levels of baseline conformity. Moreover, the mechanisms of the incentive structures in the binary choice setting of Experiment 1 (as analysed in the theoretical model) generalise to the multi-dimensional choice environment of Experiment 2.

Our experiments also shed light on the mechanisms driving evaluators' selection decisions. We find that homophily is a much more powerful rule for assigning reward and punishment than salience (unless the latter is pushed to an extreme by design). This finding calls for an extension of Rubinstein's quote, recommending to deviate from the common dress code for the reason of homophily in addition to salience. As the applicant is typically uncertain about the evaluator's preferences over clothing styles, our study would imply that *in the likely event that the evaluator prefers the conventional look, the casual outfit does not significantly hurt your chances to get the job because they are small anyway. However, in the unlikely event that you meet an unconventional evaluator like Rubinstein, the casual look will clearly increase your chances.*

Focusing on strategic conformist and anti-conformist behaviour, our paper complements the broad literature across disciplines on conformist (and, rarely, anti-conformist) preferences. Yet, preferences and behaviour are related and may mutually affect each other in both directions. Indisputably, changes in attitudes may be expressed in behavioural changes, but strategic behaviour opposing one's preferences may also feed back to preference changes to resolve the cognitive dissonance resulting from their mismatch (Festinger, 1957).

The individual level aside, societal benefits and costs of conformity and anti-conformity may vary considerably across situations, determining the use of positive or negative incentives in a specific context. Conformity can be exploited for desirable economic outcomes, in particular norm compliance.[25] On the downside, conformity may be the reason for why people make irrational financial decisions, shy away from innovative practices, are susceptible to groupthink, or communicate in echo chambers or filter bubbles. Instead, anti-conformity may lead to new practices and discoveries in organisations as well as societies, break information cascades and erode archaic social conventions—but also reduce coordination and predictability of behaviour, which can be detrimental for a society. Thus, the potential of incentives for conformity and anti-conformity looms large.

Eawag, Switzerland & University of Konstanz, Germany
University of Konstanz, Germany, Thurgau Institute of Economics, Switzerland & CESifo, Germany
Santa Fe Institute, USA & Technical University of Denmark (DTU), Denmark

Additional Supporting Information may be found in the online version of this article:

Online Appendix
Replication Package

[25] When being informed that a majority of other people will conform, people are more likely to pay taxes (Bobek *et al.*, 2007; Coleman, 2007), save energy (Schultz *et al.*, 2007; Nolan *et al.*, 2008; Allcott, 2011; Allcott and Rogers, 2014), donate to a charity (Alpizar *et al.*, 2008; Smith *et al.*, 2015) and contribute to a political party (Perez-Truglia and Cruces, 2017).

References

Allcott, H. (2011). 'Social norms and energy conservation', *Journal of Public Economics*, vol. 95(9), pp. 1082–95.

Allcott, H. and Rogers, T. (2014). 'The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation', *American Economic Review*, vol. 104(10), pp. 3003–37.

Allen, V.L. (1965). 'Situational factors in conformity', *Advances in Experimental Social Psychology*, vol. 2, pp. 133–75.

Alpizar, F., Carlsson, F. and Johansson-Stenman, O. (2008). 'Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica', *Journal of Public Economics*, vol. 92(5), pp. 1047–60.

Amabile, T.M., Goldfarb, P. and Brackfield, S.C. (1990). 'Social influences on creativity: Evaluation, coaction, and surveillance', *Creativity Research Journal*, vol. 3(1), pp. 6–21.

Amini, M., Ekström, M., Ellingsen, T., Johannesson, M. and Strömsten, F. (2017). 'Does gender diversity promote nonconformity?', *Management Science*, vol. 63(4), pp. 1085–96.

Anderson, L.R. and Holt, C.A. (1997). 'Information cascades in the laboratory', *American Economic Review*, vol. 87(5), pp. 847–62.

Apesteguia, J., Huck, S. and Oechssler, J. (2007). 'Imitation—theory and experimental evidence', *Journal of Economic Theory*, vol. 136(1), pp. 217–35.

Argyle, M. (1957). 'Social pressure in public and private situations', *Journal of Abnormal and Social Psychology*, vol. 54, pp. 172–5.

Ariely, D. and Levav, J. (2000). 'Sequential choice in group settings: Taking the road less traveled and less enjoyed', *Journal of Consumer Research*, vol. 27(3), pp. 279–90.

Asch, S. (1951). 'Effects of group pressure upon the modification and distortion of judgements', in (H. Guetzkow, ed.), *Groups, Leadership, and Men*, pp. 177–90, Pittsburgh, PA: Carnegie Press.

Asch, S. (1952). *Social Psychology*, Hoboken, NJ: Prentice Hall.

Asch, S.E. (1955). 'Opinions and social pressure', *Scientific American*, vol. 193, pp. 31–5.

Baccara, M. and Yariv, L. (2013). 'Homophily in peer groups', *American Economic Journal: Microeconomics*, vol. 5(3), pp. 69–96.

Banerjee, A.V. (1992). 'A simple model of herd behavior', *The Quarterly Journal of Economics*, vol. 107(3), pp. 797–817.

Banerjee, A. and Fudenberg, D. (2004). 'Word-of-mouth learning', *Games and Economic Behavior*, vol. 46(1), pp. 1–22.

Becker, G. (1971). *The Economics of Discrimination*, 2nd edn., Chicago: University of Chicago Press.

Belloc, M. and Bowles, S. (2013). 'The persistence of inferior cultural-institutional conventions', *American Economic Review*, vol. 103(3), pp. 93–8.

Bernheim, B.D. (1994). 'A theory of conformity', *Journal of Political Economy*, vol. 102(5), pp. 841–77.

Bertrand, M. and Duflo, E. (2017). 'Field experiments on discrimination', in (A.V. Banerjee and E. Duflo, eds.), *Handbook of Economic Field Experiments*, vol. 1, pp. 309–93, Amsterdam: North-Holland.

Bikhchandani, S., Hirshleifer, D. and Welch, I. (1992). 'A theory of fads, fashion, custom, and cultural change as informational cascades', *Journal of Political Economy*, vol. 100(5), pp. 992–1026.

Bobek, D., Roberts, R. and Sweeney, J. (2007). 'The social norms of tax compliance: evidence from Australia, Singapore and the United States', *Journal of Business Ethics*, vol. 74(1), pp. 49–64.

Bock, O., Baetge, I. and Nicklisch, A. (2014). 'hroot: Hamburg registration and organization online tool', *European Economic Review*, vol. 71, pp. 117–20.

Boehm, C. (1993). 'Egalitarian behavior and reverse dominance hierarchy', *Current Anthropology*, vol. 34(3), pp. 227–54.

Boehm, C. (2000). *Hierarchy in the Forest: The Evolution of Egalitarian Behavior*, Cambridge, MA: Harvard University Press.

Bond, M.H. and Smith, P.B. (1996). 'Cross-cultural social and organizational psychology', *Annual Review of Psychology*, vol. 47(1), pp. 205–35.

Bordalo, P., Gennaioli, N. and Shleifer, A. (2022). 'Salience', *Annual Review of Economics*, vol. 14(1), pp. 521–44.

Bowles, S. and Choi, J.K. (2013). 'Coevolution of farming and private property during the early Holocene', *Proceedings of the National Academy of Sciences*, vol. 110(22), pp. 8830–5.

Boyd, R., Gintis, H. and Bowles, S. (2010). 'Coordinated punishment of defectors sustains cooperation and can proliferate when rare', *Science*, vol. 328(5978), pp. 617–20.

Boyd, R. and Richerson, P.J. (1982). 'Cultural transmission and the evolution of cooperative behavior', *Human Ecology*, vol. 10(3), pp. 325–51.

Boyd, R. and Richerson, P.J. (1985). *Culture and the Evolutionary Process*, Chicago: University of Chicago Press.

Brehm, J.W. (1966). *A Theory of Psychological Reactance*, Oxford: Academic Press.

Buss, D.M. (2003). *The Evolution of Desire: Strategies of Human Mating*, 2nd edn., New York: Basic Books.

Byrne, D.E. (1971). *The Attraction Paradigm*, New York: Academic Press.

Cialdini, R.B. and Goldstein, N.J. (2004). 'Social influence: Compliance and conformity', *Annual Review of Psychology*, vol. 55(1), pp. 591–621.

Cialdini, R.B., Wosinska, W., Barrett, D.W., Butner, J. and Gornik-Durose, M. (1999). 'Compliance with a request in two cultures: The differential influence of social proof and commitment/consistency on collectivists and individualists', *Personality and Social Psychology Bulletin*, vol. 25(10), pp. 1242–53.

Coleman, S. (2007). 'The Minnesota income tax compliance experiment: Replication of the social norms experiment', Preprint, http://dx.doi.org/10.2139/ssrn.1393292.

Corazzini, L. and Greiner, B. (2007). 'Herding, social preferences and (non-)conformity', *Economics Letters*, vol. 97(1), pp. 74–80.
Croissant, Y. (2019). *mlogit: Multinomial Logit Models*, R package version 0.4-1.
Crutchfield, R.S. (1955). 'Conformity and character', *American Psychologist*, vol. 10(5), p. 191.
Crutchfield, R.S. (1962). 'Conformity and creative thinking', in (H.E. Gruber, G. Terrell and M. Wertheimer, eds.), *Contemporary Approaches to Creative Thinking*, pp. 120–40, New York: Atherton Press.
Currarini, S., Jackson, M.O. and Pin, P. (2009). 'An economic model of friendship: Homophily, minorities, and segregation', *Econometrica*, vol. 77(4), pp. 1003–45.
Denton, K.K., Ram, Y., Liberman, U. and Feldman, M.W. (2020). 'Cultural evolution of conformity and anti-conformity', *Proceedings of the National Academy of Sciences*, vol. 117(24), pp. 13603–14.
Dvorak, F. (2023). 'Stratest: a software package for strategy frequency estimation', *Journal of the Economic Science Association*, vol. 9(2), pp. 337–49.
Efferson, C., Lalive, R., Richerson, P.J., McElreath, R. and Lubell, M. (2008). 'Conformists and mavericks: The empirics of frequency-dependent cultural transmission', *Evolution and Human Behavior*, vol. 29(1), pp. 56–64.
Ertug, G., Brennecke, J., Kovács, B. and Zou, T. (2022). 'What does homophily do? A review of the consequences of homophily', *Academy of Management Annals*, vol. 16(1), pp. 38–69.
Falk, A. and Zimmermann, F. (2017). 'Consistency as a signal of skills', *Management Science*, vol. 63(7), pp. 2197–210.
Fatas, E., Heap, S.P.H. and Arjona, D.R. (2018). 'Preference conformism: An experiment', *European Economic Review*, vol. 105, pp. 71–82.
Fehrler, S. and Hughes, N. (2018). 'How transparency kills information aggregation: Theory and experiment', *American Economic Journal: Microeconomics*, vol. 10, pp. 181–209.
Festinger, L. (1953). 'An analysis of compliant behavior', in (M. Sherif and M.O. Wilson, eds.), *Group Relations at the Crossroads*, pp. 232–56, New York: Harper.
Festinger, L. (1957). *A Theory of Cognitive Dissonance*, Stanford, CA: Stanford University Press.
Fischbacher, U. (2007). 'z-Tree: Zurich toolbox for ready-made economic experiments', *Experimental Economics*, vol. 10(2), pp. 171–8.
Fromkin, H.L. (1970). 'Effects of experimentally aroused feelings of undistinctiveness upon valuation of scarce and novel experiences', *Journal of Personality and Social Psychology*, vol. 16(3), pp. 521–9.
Goeree, J.K. and Yariv, L. (2015). 'Conformity in the lab', *Journal of the Economic Science Association*, vol. 1(1), pp. 15–28.
Goldberg, A. and Stein, S.K. (2018). 'Beyond social contagion: Associative diffusion and the emergence of cultural variation', *American Sociological Review*, vol. 83(5), pp. 897–932.
Golub, B. and Jackson, M.O. (2012). 'How homophily affects the speed of learning and best-response dynamics', *The Quarterly Journal of Economics*, vol. 127(3), pp. 1287–338.
Goncalo, J.A. and Staw, B.M. (2006). 'Individualism-collectivism and group creativity', *Organizational Behavior and Human Decision Processes*, vol. 100(1), pp. 96–109.
Greiner, B. (2015). 'Subject pool recruitment procedures: Organizing experiments with ORSEE', *Journal of the Economic Science Association*, vol. 1(1), pp. 114–25.
Griliches, Z. (1957). 'Hybrid corn: An exploration in the economics of technological change', *Econometrica*, vol. 25(4), pp. 501–22.
Griskevicius, V., Goldstein, N.J., Mortensen, C.R., Cialdini, R.B. and Kenrick, D.T. (2006). 'Going along versus going alone: When fundamental motives facilitate strategic (non)conformity', *Journal of Personality and Social Psychology*, vol. 91(2), pp. 281–94.
Guarino, A., Harmgart, H. and Huck, S. (2011). 'Aggregate information cascades', *Games and Economic Behavior*, vol. 73(1), pp. 167–85.
Hamilton, W. (1971). 'Geometry for the selfish herd', *Journal of Theoretical Biology*, vol. 31(2), pp. 295–311.
Hegde, D. and Tumlinson, J. (2014). 'Does social proximity enhance business partnerships? Theory and evidence from ethnicity's role in US venture capital', *Management Science*, vol. 60(9), pp. 2355–80.
Hertwig, R. and Ortmann, A. (2001). 'Experimental practices in economics: A methodological challenge for psychologists?', *Behavioral and Brain Sciences*, vol. 24(3), pp. 383–451.
Hewstone, M., Rubin, M. and Willis, H. (2002). 'Intergroup bias', *Annual Review of Psychology*, vol. 53(1), pp. 575–604.
Imhoff, R. and Erb, H.P. (2009). 'What motivates nonconformity? Uniqueness seeking blocks majority influence', *Personality and Social Psychology Bulletin*, vol. 35(3), pp. 309–20.
Jones, D. and Linardi, S. (2014). 'Wallflowers: Experimental evidence of an aversion to standing out', *Management Science*, vol. 60(7), pp. 1757–71.
Jones, S.R.G. (1984). *The Economics of Conformism*, Oxford: Blackwell.
Juul, J.S. and Porter, M.A. (2019). 'Hipsters on networks: How a minority group of individuals can lead to an antiestablishment majority', *Physical Review E*, vol. 99(2), 022313.
Kelman, H.C. (1961). 'Processes of opinion change', *Public Opinion Quarterly*, vol. 25, pp. 57–78.
Kets, W. and Sandroni, A. (2016). 'Challenging conformity: A case for diversity', Preprint, http://dx.doi.org/10.2139/ssrn.2871490.
Kets, W. and Sandroni, A. (2021). 'A theory of strategic uncertainty and cultural diversity', *The Review of Economic Studies*, vol. 88(1), pp. 287–333.

Kim, H. and Markus, H.R. (1999). 'Deviance or uniqueness, harmony or conformity? A cultural analysis', *Journal of Personality and Social Psychology*, vol. 77(4), pp. 785–800.
Krupka, E.L. and Weber, R.A. (2013). 'Identifying social norms using coordination games: Why does dictator game sharing vary?', *Journal of the European Economic Association*, vol. 11(3), pp. 495–524.
Lazarsfeld, P.F. and Merton, R.K. (1954). 'Friendship as a social process: A substantive and methodological analysis', in (M. Berger, T. Abel and H. Charles, eds.), *Freedom and Control in Modern Society*, pp. 18–66, New York: Van Nostrand.
Lee, I.H. (1993). 'On the convergence of informational cascades', *Journal of Economic Theory*, vol. 61(2), pp. 395–411.
Lee, I.H. (1998). 'Market crashes and informational avalanches', *The Review of Economic Studies*, vol. 65(4), pp. 741–59.
Loomes, G., Starmer, C. and Sugden, R. (1991). 'Observing violations of transitivity by experimental methods', *Econometrica*, vol. 59, pp. 425–39.
Lynn, M. and Harris, J. (1997). 'Individual differences in the pursuit of self-uniqueness through consumption', *Journal of Applied Social Psychology*, vol. 27(21), pp. 1861–83.
Mahdi, N.Q. (1986). 'Pukhtunwali: Ostracism and honor among the Pathan Hill tribes', *Ethology and Sociobiology*, vol. 7(3–4), pp. 295–304.
Mäkelä, K., Björkman, I. and Ehrnrooth, M. (2010). 'How do MNCs establish their talent pools? Influences on individuals' likelihood of being labeled as talent', *Journal of World Business*, vol. 45(2), pp. 134–42.
March, J.G. (1991). 'Exploration and exploitation in organizational learning', *Organization Science*, vol. 2(1), pp. 71–87.
Matusik, S.F., George, J.M. and Heeley, M.B. (2008). 'Values and judgment under uncertainty: Evidence from venture capitalist assessments of founders', *Strategic Entrepreneurship Journal*, vol. 2(2), pp. 95–115.
McElreath, R., Bell, A.V., Efferson, C., Lubell, M., Richerson, P.J. and Waring, T. (2008). 'Beyond existence and aiming outside the laboratory: Estimating frequency-dependent and pay-off-biased social learning strategies', *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 363(1509), pp. 3515–28.
McFadden, D. (1974). 'Conditional logit analysis of qualitative choice behavior', in (P. Banerjee, ed.), *Fontiers in Econometrics*, pp. 105–42, New York: Academic Press.
McPherson, M., Smith-Lovin, L. and Cook, J.M. (2001). 'Birds of a feather: Homophily in social networks', *Annual Review of Sociology*, vol. 27(1), pp. 415–44.
Mehta, J., Starmer, C. and Sugden, R. (1994a). 'Focal points in pure coordination games: An experimental investigation', *Theory and Decision*, vol. 36(2), pp. 163–85.
Mehta, J., Starmer, C. and Sugden, R. (1994b). 'The nature of salience: An experimental investigation of pure coordination games', *The American Economic Review*, vol. 84(3), pp. 658–73.
Nail, P.R. (1986). 'Toward an integration of some models and theories of social response', *Psychological Bulletin*, vol. 100(2), pp. 190–206.
Nail, P.R., Di Domenico, S.I. and MacDonald, G. (2013). 'Proposal of a double diamond model of social response', *Review of General Psychology*, vol. 17(1), pp. 1–19.
Nail, P.R. and Van Leeuwen, M.D. (1993). 'An analysis and restructuring of the diamond model of social response', *Personality and Social Psychology Bulletin*, vol. 19(1), pp. 106–16.
Nolan, J.M., Schultz, P.W., Cialdini, R.B., Goldstein, N.J. and Griskevicius, V. (2008). 'Normative social influence is underdetected', *Personality and Social Psychology Bulletin*, vol. 34(7), pp. 913–23.
Nyczka, P., Byrka, K., Nail, P.R. and Sznajd-Weron, K. (2018). 'Conformity in numbers—does criticality in social responses exist?', *PLoS ONE*, vol. 13(12), e0209620.
Nyczka, P. and Sznajd-Weron, K. (2013). 'Anti-conformity or independence? Insights from statistical physics', *Journal of Statistical Physics*, vol. 151(1), pp. 174–202.
Opper, S., Nee, V. and Brehm, S. (2015). 'Homophily in the career mobility of China's political elite', *Social Science Research*, vol. 54, pp. 332–52.
Perez-Truglia, R. and Cruces, G. (2017). 'Partisan interactions: Evidence from a field experiment in the United States', *Journal of Political Economy*, vol. 125(4), pp. 1208–43.
Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M.W., Fogarty, L., Ghirlanda, S., Lillicrap, T. and Laland, K.N. (2010). 'Why copy others? Insights from the social learning strategies tournament', *Science*, vol. 328(5975), pp. 208–13.
Riach, P.A. and Rich, J. (2002). 'Field experiments of discrimination in the market place', Economic Journal, vol. 112(483), pp. F480–518.
Robin, S., Rusinowska, A. and Villeval, M.C. (2014). 'Ingratiation: Experimental evidence', *European Economic Review*, vol. 66, pp. 16–38.
Rubinstein, A. (2013). '10 Q&A: Experienced advice for "lost" graduate students in economics', *The Journal of Economic Education*, vol. 44(3), pp. 193–6.
Sakha, S. and Grohmann, A. (2016). 'The effect of peer observation on consumption choices: Experimental evidence', Preprint, http://dx.doi.org/10.2139/ssrn.2797074.
Schaerf, T.M., Dillingham, P.W. and Ward, A.J.W. (2017). 'The effects of external cues on individual and collective behavior of shoaling fish', *Science Advances*, vol. 3(6), doi: 10.1126/sciadv.1603201.
Schlag, K.H. (1998). 'Why imitate, and if so, how?: A boundedly rational approach to multi-armed bandits', *Journal of Economic Theory*, vol. 78(1), pp. 130–56.
Schlag, K.H. (1999). 'Which one should I imitate?', *Journal of Mathematical Economics*, vol. 31(4), pp. 493–522.

Schmelz, K. (2011). 'E-nstructions: A tool for electronic instructions in laboratory experiments', Jena Economic Research Paper 2011-008.
Schultz, P.W., Nolan, J.M., Cialdini, R.B., Goldstein, N.J. and Griskevicius, V. (2007). 'The constructive, destructive, and reconstructive power of social norms', *Psychological Science*, vol. 18(5), pp. 429–34.
Schumpeter, J.A. (1934). *The Theory of Economic Development*, Cambridge, MA: Harvard University Press.
Schwarz, G. (1978). 'Estimating the dimension of a model', *Annals of Statistics*, vol. 6(2), pp. 461–4.
Selten, R. (1967). 'Die strategiemethode zur erforschung des eingeschränkt rationalen verhaltens im rahmen eines oligopolexperiments', in (H. Sauermann, ed.), *Beiträge zur experimentellen Wirtschaftsforschung*, pp. 136–68, Tübingen: Mohr.
Shalley, C.E. and Perry-Smith, J.E. (2001). 'Effects of social-psychological factors on creative performance: The role of informational and controlling expected evaluation and modeling experience', *Organizational Behavior and Human Decision Processes*, vol. 84(1), pp. 1–22.
Shane, S.A. (1992). 'Why do some societies invent more than others', *Journal of Business Venturing*, vol. 7(1), pp. 29–46.
Siedlecki, P., Szwabinski, J. and Tomasz, W. (2016). 'The interplay between conformity and anticonformity and its polarizing effect on society', *Journal of Artificial Societies and Social Simulation*, vol. 19(4), doi: 10.18564/jasss.3203.
Simon, H.A. (1955). 'A behavioral model of rational choice', *The Quarterly Journal of Economics*, vol. 69(1), pp. 99–118.
Simpson, J.A., Gangestad, S.W., Christensen, P.N. and Leck, K. (1999). 'Fluctuating asymmetry, sociosexuality, and intrasexual competitive tactics', *Journal of Personality and Social Psychology*, vol. 76(1), pp. 159–72.
Smith, L. and Sorensen, P. (2000). 'Pathological outcomes of observational learning', *Econometrica*, vol. 68(2), pp. 371–98.
Smith, S., Windmeijer, F. and Wright, E. (2015). 'Peer effects in charitable giving: Evidence from the (running) field', Economic Journal, vol. 125(585), pp. 1053–71.
Soll, J.B. and Larrick, R.P. (2009). 'Strategies for revising judgment: How (and how well) people use others' opinions', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 35(3), p. 780.
Sosna, M.M.G., Twomey, C.R., Bak-Coleman, J., Poel, W., Daniels, B.C., Romanczuk, P. and Couzin, I.D. (2019). 'Individual and collective encoding of risk in animal groups', *Proceedings of the National Academy of Sciences*, vol. 116(41), pp. 20556–61.
Tajfel, H. (1970). 'Experiments in intergroup discrimination', *Scientific American*, vol. 223(5), pp. 96–103.
Tibbetts, E.A. and Dale, J. (2007). 'Individual recognition: It is good to be different', *Trends in Ecology & Evolution*, vol. 22(10), pp. 529–37.
Touboul, J. (2019). 'The hipster effect: When anti-conformists all look the same', *Discrete & Continuous Dynamical Systems—B*, vol. 24(8), pp. 4379–415.
Turner, J.C., Brown, R.J. and Tajfel, H. (1979). 'Social comparison and group interest in in-group favouritism', *European Journal of Social Psychology*, vol. 9(2), pp. 187–204.
Tversky, A. (1969). 'Intransitivity of preferences', *Psychological Review*, vol. 76(1), pp. 31–48.
Vega-Redondo, F. (1997). 'The evolution of Walrasian behavior', *Econometrica*, vol. 65(2), pp. 375–84.
Vine, I. (1971). 'Risk of visual detection and pursuit by a predator and the selective advantage of flocking behaviour', *Journal of Theoretical Biology*, vol. 30(2), pp. 405–22.
Viscido, S.V. and Wethey, D.S. (2002). 'Quantitative analysis of fiddler crab flock movement: Evidence for selfish herd behaviour', *Animal Behaviour*, vol. 63(4), pp. 735–41.
Vives, X. (1997). 'Learning from others: A welfare analysis', *Games and Economic Behavior*, vol. 20(2), pp. 177–200.
Wiessner, P. (2002). 'Hunting, healing, and hxaro exchange—a long-term perspective on !Kung (Ju/'hoansi) large-game hunting', *Evolution and Human Behavior*, vol. 23(6), pp. 407–36.
Willis, R.H. (1963). 'Two dimensions of conformity-nonconformity', *Sociometry*, vol. 26(4), pp. 499–513.
Willis, R.H. (1965). 'Conformity, independence, and anti-conformity', *Human Relations*, vol. 18(4), pp. 373–88.
Willis, R.H. and Levine, J.M. (1976). 'Interpersonal influence and conformity', in (B. Seidenberg and A. Snadowsky, eds.), *Social Psychology: An introduction*, pp. 309–41, New York: Free Press.
Wright, D.B., London, K. and Waechter, M. (2009). 'Social anxiety moderates memory conformity in adolescents', *Applied Cognitive Psychology*, vol. 24(7), pp. 1034–45.
Yamagishi, T., Hashimoto, H. and Schug, J. (2008). 'Preferences versus strategies as explanations for culture-specific behavior', *Psychological Science*, vol. 19(6), pp. 579–84.
Young, H.P. (2009). 'Innovation diffusion in heterogeneous populations: Contagion, social influence, and social learning', *American Economic Review*, vol. 99(5), pp. 1899–924.

The Economic Journal, **135** (*February*), 556–583 https://doi.org/10.1093/ej/ueae085 © The Author(s) 2024. Published by Oxford University Press on behalf of Royal Economic Society.
Advance Access Publication Date: 25 September 2024