

Crowding-out the in-group bias: a nationalist policy paradox?

Shaun P. Hargreaves Heap ^a, Eugenio Levi ^b, Abhijit Ramalingam ^c

^a Department of Political Economy, King's College London, The Strand, London WC2, UK,
s.hargreavesheap@kcl.ac.uk, Tel: +44-20-7848-1689

^b Department of Public Economics, Masaryk University, Lipová 507/41a, Brno, 602 00, CZ,
eugenio.levi@econ.muni.cz

^c Department of Economics, Walker College of Business, Appalachian State University,
Peacock Hall, Boone, NC 28608, USA, ramalingama@appstate.edu, Tel: +1-828-262-2418

October 2020

Abstract

Using a dictator game experiment, we investigate if a policy of introducing material incentives to favour one's own group members will be effective in raising the in-group bias in behaviour. It is not: the introduction of the material incentives in our experiment crowds-out the in-group bias in our subjects' social preferences. Specifically, we find evidence that is consistent with the social identification with own group members weakening through the introduction of material incentives towards the in-group bias. This result potentially creates a nationalist policy paradox whereby policies like tariffs and discriminatory employment regulations designed to encourage materially the employment of home rather than foreign workers will, on the evidence of this experiment, weaken individuals' preferences for favouring home over foreign workers.

JEL Codes: C72, C91, D31, D63, D91, J70, Z18

Keywords: experiment, dictator game, social identification, in-group bias, incentives, crowding-out

People are frequently nicer to members of their own group than those who belong to a different one. This in-group bias in pro-sociality has, for example, been frequently observed in social psychology and economics (e.g. see Chen and Li, 2009, and Hargreaves Heap and Zizzo, 2009, for experimental evidence). In this paper, we examine with an experiment whether this revealed in-group bias in social preferences is crowded-out by the introduction of material incentives designed to encourage in-group biased behaviour.

In general, the possible crowding-out of social preferences is important because it can affect the efficacy of policy interventions that turn on tweaking material incentives in favour of pro-social behaviour: the weakening of social preferences tends to offset the effect on behaviour of the change in the material incentives. In our particular case, a crowding-out of the in-group bias in social preferences through the introduction of pro-in-group material incentives would have a paradoxical policy implication. We call it the nationalist policy paradox. This is because common nationalist policies, like tariffs and tougher employment regulations for foreign workers, that materially encourage the employment of home rather than foreign workers, would, with crowding-out, paradoxically mean that the motivating belief or social preference for such policies of treating home workers better than foreign ones would actually become weaker.

The background to this question is a large literature on the crowding-out of social preferences when material incentives designed to encourage pro-social behaviours are introduced (see Bowles and Polania-Reyes, 2012, for a survey). Gneezy and Rustichini (2000), hereafter G&R, famously illustrate this possibility and the associated policy concern. They report on an experiment where a fine, introduced to deter late pick-ups at day care nurseries, backfires spectacularly because the numbers of late pick-ups actually increases after the introduction of the fine. We qualitatively replicate the G&R experiment, but, in a different laboratory setting, to test for the possible crowding-out of the in-group biased character of social preferences.

To our knowledge, we are the first to test for this possibility and its associated implication of a nationalist policy paradox. This is one of our contributions and its relevance stretches beyond that of nationalist policies. Companies or teams, for example, that compete with each other might naturally wish to encourage their employees/team members to behave more nicely and more cooperatively with each other than with their competitors' employees/team members. Would a strengthening of the material incentive towards being especially nice to own group

members be an effective way of encouraging this difference in behaviour or should they fear crowding-out?

There are also examples where policy interventions could take the opposite form because the in-group bias appears to be unwarranted discrimination. For example, there is no good reason for a doctor or lawyer to care more about a patient or a client simply because they belong to the same group as themselves. So, the question arises: would a tweak in the material incentives, this time against the in-group bias, be effective in reducing the in-group bias in behaviour? Or might there be some countervailing crowding-out, in this instance, of the social preference for equal treatment that will tend to offset the effect of the change in material incentives? (In an ancillary experiment that we report in the appendix, we test, in manner analogous to the crowding-out of the in-group bias in social preferences in our main experiment, whether there is also crowding-out of the equal treatment social preferences when material incentives are introduced to discourage the in-group bias.)

The experimental methodology enables us to identify whether there is crowding-out. It also allows us to test a particular explanation of the in-group bias and its possible crowding-out. This is our second contribution: we test whether the in-group bias arises because people socially identify more strongly with own group members than others; and we test whether, if there is crowding-out, this can be explained because own group social identification weakens with the introduction of the material incentives. Social identification theory provides a plausible explanation of the in-group bias (e.g., see Tajfel and Turner, 1979), but it is not the only one. Out-group hate is another possible explanation of the bias and, in so far as there is crowding out, it could also arise from a weakening in out-group hate. This difference in the possible origin of the in-group bias that we test can also be expressed slightly differently: does the in-group bias arise from positive or negative discrimination (see Hargreaves Heap and Zizzo, 2009)?

Our second contribution in this respect is also potentially important, partly because social identification theory has become an increasingly popular explanatory vehicle in economics (see Akerlof and Kranton, 2000 & 2005, and Shayo, 2020). It is also important because in so far as there is crowding-out and it can be connected to weakening social identification, then it points to a more general conclusion: the preferences that are revealed through the influence of social identification on behaviour are not fixed. That is, they cannot be taken to be exogenous. This matters because preferences are often regarded as a bedrock in economics. For example, Stigler

and Becker (1977) famously suggest that ‘de gustibus non est disputandum’ and Lucas (1976) notably argues for and establishes a programme in macroeconomics that is based on individual preferences precisely because they are presumed to be stable.

In the experiment, our subjects make dictator decisions in three phases either in a baseline control where there is no group affiliation or in group treatments where subjects are randomly assigned to either a Yellow or Green group. In the group treatments, each subject makes two dictator decisions in each phase: one where the co-player comes from own group and the other where co-player belongs to the other group; and the group affiliations are common knowledge.¹

Our background assumption is that individuals decide how much to allocate to their co-player by weighing their selfish preference for own pay-offs against their social preference for the pay-offs of the co-player. In the first phase of these dictator decisions this is the only consideration because the dictator simply has an endowment and makes the allocation to the co-player (there is no policy of a fine or a subsidy to provide an extra material incentive either towards or away from an allocation to the co-player). We further conjecture from social identification theory that in the group treatments the social preference weight attached to the co-player’s pay-offs is higher when the other person belongs to the same group than when there are no group affiliations; whereas the weight attached to a co-player from the other group is no higher in the group treatments than when there are no group affiliations in the baseline control.

Aggregate behaviour in the first phase is consistent with this prediction, but we find individual differences. Roughly half our subjects behave in this way and reveal the in-group bias in social preferences and half either make no such distinction by giving the same amount to both types of co-player or give more to a co-player from the other group. Although the balance between these two groups is somewhat different in our experiment, this is not unlike the G&R first phase because they start from a position where some people reveal a social preference for ‘good’ behaviour with timely pick-ups and others reveal with late-pick-ups either no such social preference or, indeed, a social preference for ‘bad’ behaviour.

In the second phase, G&R introduce a fine on ‘bad’ behaviour in some day care centres and not others. We do the same in phase 2. We have one Group treatment (Group-Fine) where the subjects in phase 2 are fined if they do not exhibit an in-group bias in their allocation in the second phase and another group treatment (Group) where there is no fine. In other words, our

¹ Thus, the group affiliations are artificial and minimal and so provide a ‘tough’ test of social identification in the sense of Popper.

in-group biased social preference is analogous to the ‘good’ behaviour social preference in G&R and we attempt to encourage the behaviour associated with this preference by fining those who do not behave in this manner in Group-Fine, just as G&R do.² Despite the fine on those who do not exhibit the in-group bias, the aggregate in-group bias does not change in the phase 2 of Group-Fine as compared with either that in the first phase or when compared with the phase 2 in-group bias in Group (the group treatment where there is no fine). Thus, although the policy does not spectacularly backfire in the way of G&R, the fine policy is nevertheless ineffective in our experiment and this points to the existence of crowding-out.

This aggregate evidence of crowding-out is what G&R present in favour of crowding-out. Our laboratory design, however, improves over the G&R test for crowding-out because we can also test for crowding-out at the individual level and the possible mechanism behind it. In particular, we find that those who reveal the in-group bias social preference in the first phase of Group-Fine reduce the extent of their bias in the second phase. This is important because those who have revealed a social preference for the in-group bias in the first phase, have no material reason to adjust their behaviour in the second phase when the fine is introduced. The reduction in their in-group bias can only have arisen because their social preference for the in-group bias diminished: i.e. it was crowded-out. Furthermore, and this is the part that explicitly refers to social identification mechanism, we find that the reduction in the in-group bias occurs because the allocation to own group members falls. The weight given to the pay-offs of a co-player from own group falls and this is consistent with the fine actually weakening the dictator’s social identification with own group members.

Finally, in the third phase in Group-Fine, like G&R, we remove the fine and examine whether the crowding-out in phase 2 persists. Again, we can test for persistence in the aggregate data like G&R and, in addition, through individual level data that also allows us to test the social identification mechanism. Unlike G&R, the crowding-out does not persist in our experiment.

In the next section, we define in a dictator decision the in-group bias, an in-group bias in social preferences and their crowding out that we will test and we develop the hypotheses we use to test the possible role of social identification theory in explaining this social preference bias and its change. Section 2 explains the experimental design and Section 3 gives the results. Section 4 concludes.

² Our electronic appendix describes a complementary experiment where we instead fine in-group bias behaviour so as to discourage this kind of behaviour. We focus in the main part of the paper on the fine to encourage in-group biased behaviour for technical reasons that we explain later.

1. Theory and hypotheses

We ask our subjects in the group treatment to make dictator decisions with co-players who either belong to the same group or the other group. We define the possible varieties of biased and non-biased behaviour by the relations between allocations to the co-player who is a member of the same group ($= CP(own)$) and the co-player who belongs to the other group ($= CP(other)$). These supply the tests for whether there is an in-group bias and whether it changes.

Definition: In-group biased behaviour (IGB) arises when $CP(own) > CP(other)$ and its extent is measured by $CP(own) - CP(other)$.

Definition: Equal treatment behaviour (EQB) arises when $CP(own) = CP(other)$.

Definition: Out-group biased behaviour (OGB) arises when $CP(own) < CP(other)$ and its extent is measured by $CP(own) - CP(other)$.

We choose the gap between own and other allocations as the index of the in-group bias, but recognise that a ratio measure could have been used. Accordingly, the Appendix gives the corresponding results for the ratio measure. There are no qualitative differences.

We assume in general that individuals value their own pay-off (OP) and (possibly) their co-player's pay-off (CP) as in (1).

$$U = f(OP, CP) \quad (1)$$

In the first phase dictator decision, an individual decides how to divide a sum X between OP and CP . This is the constraint on maximising (1). Since the relative 'price' of OP in terms of CP is 1 in this constraint, it follows that utility maximisation will be achieved when the ratio of marginal utilities from OP and CP is equal to this relative price of 1. The chosen allocation OP/CP is thus given by the elasticity of substitution between OP and CP in (1). The smaller the elasticity (i.e. the larger the % change in CP is required to compensate for a unit % change in OP), the bigger is the share of OP relative to CP .

As an illustration consider a Cobb-Douglas utility function as in (1'), where 'a' and 'b' are the weights given respectively to each type of pay-off in the individual's utility function, and A is a constant. In effect, this follows the Charness and Rabin (2000) representation of preferences when they test for the character of social preferences revealed in dictator like decisions. They consider discrete choices between pairs of allocation and so can use a linear utility function in

own and co-player pay-offs. As we have a range of options between 0% and 100% of X , this linearity would produce corner solutions and to avoid this we assume log-linear preferences.

$$U = A * OP^a * CP^b \quad (1')$$

Maximising (1) subject to the constraint $OP + CP = X$ yields the following:

$$OP = a * X / (a + b)$$

$$CP = b * X / (a + b) \quad (2)$$

In the simple dictator game above, we note that there is no material incentive in this utility maximisation dictator decision to treat co-players differently on the basis of their group membership because a one-unit allocation to a co-player costs that individual one unit in terms of OP whether the co-player comes from own or the other group. Thus, in so far as $CP(own) > CP(other)$ (i.e. IGB is observed), it reveals in-group biased social preferences (IGBSP). For example, in the Cobb-Douglas illustration ' $b(own) > b(other)$ '. By the same reasoning in this simple dictator decision, EQB reveals an equal treatment social preferences (EQTSP) and OGB reveals an out-group biased social preferences (OGBSP).

In this way the relation between $CP(own)$ and $CP(other)$ tells us whether IGBSP, EQTSP or OGBSP are revealed by subjects when they make the simple dictator decision above.

We assume for the purpose of testing social identification theory that it predicts that individuals who identify more closely with a group weigh co-player's pay-offs from that group more highly than they do co-player's from groups they identify with less closely. Thus for example, ' $b(own) > b(other)$ ', ' b (the value in the no – group control) in the Cobb-Douglas representation of preferences. Such social identification, together with (2) implies $CP(own) > CP(other)$, $CP(when\ there\ are\ no\ groups)$.³ In so far as social identification is weak or does not apply, then ' $b(own) \cong b(other) \cong b$ (the value in the no – group control) and $CP(own) \cong CP(other) \cong CP(when\ there\ are\ no\ groups)$.

Social identification theory provides one reason why IGBSP might be revealed in behaviour in the simple dictator decision, but it is not the only possible cause. An alternative explanation of the bias in behaviour is that the introduction of explicit groups triggers out-group hate. In this case, we assume $CP(other)$ falls relative to CP when there are no groups (i.e. ' $b(other)$ ' falls

³ In the general case, we assume social identification theory predicts that the elasticity of substitution between OP and CP in an individual's utility function is higher when the co-player is from own group: i.e., it requires a smaller % change in an own group co-player's pay-offs to compensate for a unit % change in own pay-offs.

relative to 'b'(where there are no groups). Further since there is no reason to suppose $CP(own)$ is different to CP on this account (i.e., ' $b(own)$ ' is any different to 'b'), a gap is opened up between $CP(own)$ and $CP(other)$ because $CP(other)$ falls (i.e., ' $b(own)$ ' > ' $b(other)$ ' because $b(other)$ falls).

Thus our basic test of social identification versus out-group hate in the explanation of IGB is whether the gap between $CP(own)$ and $CP(other)$ opens up because $CP(own) > CP$ or because $CP(other) < CP$. Thus an alternative way of expressing this difference is whether the in-group bias arises from positive discrimination in favour of own group members or negative discrimination against out-group members.

It is, of course, possible that subjects feel some identification with the group as a whole when there are no explicit group affiliations in our Control---so CP may reflect some social identification. This is less likely in our online experiment than in laboratory ones. Nevertheless, such a whole group identification with everyone in the experiment, if it exists, cannot be as strong as the identification with own group when there is explicit assignment to either a Yellow or Green group: thus, with group identification, $CP(own)$ will be greater than CP (i.e., $b(own)$ will be greater than 'b'), and how much greater depends on whether there is any whole group identification when there are no explicit groups. In so far as there was any whole group identification supporting CP , then this means that social identification theory might also predict that CP will be greater than $CP(other)$ (i.e. 'b' could be higher than ' $b(other)$ ') and this, of course, is what the out-group hate hypothesis predicts. So in these circumstances what distinguishes the social identification account from out-group hate is that $CP(own)$ exceeds CP (i.e., ' $b(own)$ ' will be greater than 'b'). H1 follows.

H1 (social identification and in-group bias): $CP(own)$ is greater than $CP(other)$ because relative to CP when there are no groups, the introduction of explicit groups leads $CP(own)$ to rise and $CP(other)$ does not rise.

Now, let us consider how material incentives could influence these social preferences. There is a large social psychology literature following Deci (1975) arguing that the 'intrinsic' reasons for taking an action can be crowded-out by the introduction of 'extrinsic' reasons to take that action. 'Intrinsic' reasons have often been taken in economics to mean having a preference for that action (or its outcome) and the 'extrinsic' reasons for action come from material incentives towards an action (e.g. see Frey, 1997). This literature predicts that the introduction of a material incentive towards a behaviour may so crowd-out the intrinsic reasons for the action

that the incentive has no or possibly the opposite effect on behaviour in the aggregate. This is the version of crowding-out that G&R test and we do the same by introducing a fine in the phase 2 dictator decisions on those who do not exhibit IGB. H2 follows as an analogous test to G&R of crowding-out in our experiment.

H2 (aggregate crowding-out): The introduction of the fine in phase 2 designed to encourage IGB either has no effect on the IGB in the aggregate or a negative effect (i.e., IGB falls).

We are also able to test for crowding-out at the individual level. Consider formally a second dictator decision problem where a fine (F) is introduced on any individual who does not reveal IGB. The fine creates a new constraint for the maximisation problem, given by (3).

$$\begin{aligned} OP &= X - CP && \text{if } CP(\text{own}) > CP(\text{other}) \\ OP &= X - CP - F && \text{if } CP(\text{own}) \leq CP(\text{other}) \end{aligned} \quad (3)$$

We note that, for those who revealed IGBSP in the first phase dictator decisions, this new constraint is not binding on the utility maximizing decision. There is a change in material incentives but that change does not materially impinge on decision makers who revealed IGBSP in phase 1. Thus, the only reason for subjects who reveal IGB behaviour in phase 1 to change their IGB behaviour in phase 2 is if the IGBSP changes in phase 2. H3 follows as a test that it occurs in our experiment.⁴

H3 (individual crowding-out): Those who reveal IGBSP in the phase 1 dictator decision reveal lower IGB behaviour in the phase 2 than in phase 1: i.e., their $CP(\text{own}) - CP(\text{other})$ falls in the second phase compared with the first.

The alternative hypotheses for this set of individuals with IGBSP in phase 1 are either that there is no change in $CP(\text{own}) - CP(\text{other})$ or that there $CP(\text{own}) - CP(\text{other})$ rises. We call the latter crowding-in.

Although this reverse possibility has not been theorised in the same way as crowding-out, there is evidence of it in the empirical literature (see Bowles and Polania-Reyes, 2012). It is also not

⁴ In the complementary experiment where the fine is levied on IGB behaviour, the analogous test would be that those who initially revealed EQTSP and OGBSP (and so were unaffected materially by the fine) nevertheless reduced their out-group dictator allocation. This is a weaker test than the one above because, unlike IGBSP subjects above, the EQTSP subjects in the complementary experiment have no margin of adjustment in their behaviour to reveal such crowding-out while maintaining EQTSP. Crowding-out would only potentially register among OGBSP subjects and they are small in number.

difficult to see why it might occur. When a policy of encouraging a particular behaviour is introduced through tweaking the material incentives, it is possible that this public material endorsement of the behaviour encourages people to re-evaluate positively the ‘intrinsic’ reasons that they have for engaging in such actions.

If there is crowding-out then it could be explained by the crowding-out of social identification with own group (or equivalently a weakening of positive discrimination). In this case, the reason $CP(own) - CP(other)$ falls in the second phase is because $CP(own)$ falls in the second phase (it gets closer to CP with the weakening of own group social identification). The contrasting explanation of IGB behaviour that it comes through the triggering of out-group hate would instead have any crowding-out explained by the fall in out-group hate (or equivalently a weakening of negative discrimination): i.e., $CP(own) - CP(other)$ falls because $CP(other)$ rises in the second phase (e.g., in the Cobb-Douglas illustration $db(own)/dF < 0$). H4 follows.

H4 (individual crowding-out due to weakened social identification): If those who reveal IGBSP in the first phase also reveal a lower IGB ($CP(own) - CP(other)$) in the second phase than the first, it is because $CP(own)$ falls in the second phase.

We also consider a possible kind of crowding out/in of IGBSP that might arise with individuals who reveal EQTSP in the first phase. The fine in the second phase dictator decisions does affect their utility maximising decision, it creates a material incentive to move towards IGB behaviour. With a Cobb-Douglas utility function, they should marginally adjust $CP(own)$ up and/or $CP(other)$ down so that the utility cost of IGB behaviour is minimised by adjusting on both sides of EQB. Thus, in so far as EQTSP individuals in phase 1 reveal larger IGB behaviour in phase 2 than these marginal adjustments, it suggests that they have to some degree gained a social preference for IGB (i.e., IGBSP has been crowded-in).

H5 (individual crowding-in of IGBSP among the EQTSP): Those who reveal EQTSP in the first phase and adjust to the fine with IGB behaviour in the second phase, do so with non-marginal changes to $CP(own)$ and/or $CP(other)$

To preserve the comparison with G&R, we are finally interested in whether any crowding-out/in of IGBSP persists when the material incentives to IGB behaviour in Phase 2 are removed. Thus, in Phase 3 of the dictator decisions, the fine is removed and the dictator decisions are formally the same as in phase 1. Our tests for persistence follow in natural

extension naturally by comparing $CP(own) - CP(other)$ in phase 3 with that in phase 2 and phase 1.

2. Experimental design and procedures

At the beginning of the experiment, each subject received a separate one-time lump sum endowment of 50 tokens. They then made decisions in three Phases.

2.1 Dictator decisions

In each Phase, all subjects independently made decisions in a dictator game. Each subject decided how to split 80 tokens between him/herself and an anonymous subject in the study. The recipient had no say in the allocation. Before making their decisions, dictators were informed that both they and the recipient had an endowment of 50 tokens. The Nash equilibrium is for dictators to allocate 0 tokens to recipients, and keep all 80 tokens for themselves. In the absence of distributional concerns, any allocation of tokens between the two is efficient.

Before making decisions in a Phase, subjects were informed that they would be matched with a randomly chosen participant in the study, and that either their decision or that of the matched coparticipant would be implemented. This payment procedure made it clear that there was an equal chance of being a dictator or a recipient in the Phase. Therefore, it made decisions incentive compatible, i.e., subjects had every incentive to take each decision seriously.⁵

2.2 Treatments

We ran three main treatments. Treatments varied in whether or not subjects were assigned to groups, and whether dictators received incentives to favour members of their own group.

In BASELINE, subjects were not assigned to any groups and did not receive any additional incentives. In each Phase, dictators made one allocation decision where the recipient was a randomly chosen participant in the same treatment. All three Phases were identical.

In Group, subjects were randomly assigned to either a **YELLOW** or a **GREEN** group, and informed of the group assignment at the beginning of Phase 1. In each Phase, dictators made

⁵ Prior to the main experiment, all subjects independently performed a real effort task for three minutes. The task involved converting a randomly generated three-letter “word” into a numeric string (Erkal et al., 2011). Subjects were paid 3 tokens for every correct code. They received no feedback until the end of the experiment. This task was completely independent of the dictator game.

two allocation decisions: one where the recipient belonged to the same group, and one where the recipient belonged to the other group. All three Phases were identical.

In Group-Fine, subjects were once again randomly assigned to groups and made two decisions in each Phase as in Group and in phase 1 the decision is identical to that in Group. Group-Fine differs in the phase 2 dictator decisions: earnings in Phase 2 were subject to a possible adjustment. In particular, if a dictator’s decision was chosen as the allocation relevant for earnings in Phase 2, then the dictator’s earnings for the Phase were reduced by 10 tokens if he/she allocated strictly fewer tokens to the recipient from his/her own group than to a recipient from the other group. Equal allocations were also penalised. Thus, there was an incentive to favour, i.e., allocate more to, a recipient from the dictator’s own group. If the matched coparticipant’s decision was chosen for implementation, then the recipient’s earnings were not adjusted. Phase 3 like Phase 1 was identical to those Phases in Group, and earnings in these Phases were calculated as before with no adjustments. Table 1 summarises our treatments.⁶

Table 1. Summary of treatments

| Treatment | Groups? | # decisions per Phase | Earnings reduction in Phase 2? | # subjects | | Total |
|--|---------|--------------------------|-----------------------------------|------------|-------|-------|
| | | | | Yellow | Green | |
| BASELINE | No | 1 | No | 38 | | 38 |
| Group | Yes | 2 | No | 37 | 34 | 71 |
| Group-Fine | Yes | 2 | Yes, if no in-group bias | 39 | 39 | 78 |
| Earnings were adjusted <i>only if</i> the dictator’s choice was chosen for implementation. | | | | Total | | 187 |

2.3 Procedures

The experiment was conducted over two sessions using the online platform Prolific which gave us access to volunteer adult subjects from a number of countries.⁷ Upon agreeing to participate in the study advertised on Prolific, subjects were directed to a website that hosted our experiment. Subjects first read a consent statement and, if they agreed, were then presented with instructions for the experiment (available in Appendix A in the Electronic Supplementary

⁶ As mentioned earlier, we ran an additional complementary treatment. Group-FineProEqual, was procedurally the same as Group-Fine, but differed in the earnings adjustment in Phase 2. In this treatment, dictators were given an economic incentive to *not* favour recipients from their own group. Earnings were reduced by 10 tokens if a dictator’s decision was chosen as the relevant one for payment and if he/she had allocated more to a recipient who belonged to their own group. Here, equal allocations were *not* penalised. Phase 1 and Phase 3 were the same as in Group. A total of 79 (41 Yellow and 38 Green) participated in this treatment.

⁷ We conducted multiple sessions to minimise the chances of server overload during a session and to avoid the whole session crashing. The two sessions were conducted one after the other on the same day. We ran a third session where all subjects were assigned to Group-FineProEqual.

Material). Subjects were randomly assigned to one of the three treatments as they signed up to participate. They then completed the experiment on their own devices at their own pace.⁸ The experiment was programmed in oTree (Chen et al., 2016).

Subjects received no feedback during the experiment. Subjects were paid a flat participation fee of USD 1.50 upon completion of the experiment. Within the next two days, they were paid their earnings from each Phase of the experiment. Token earnings were converted to cash at the rate of 200 tokens to USD 1. The average participant took about 12 minutes to complete the experiment and received an additional USD 1.10. The total average payment was USD 2.60, which translates to USD 13 as an hourly rate.

3. Results

Table 2 gives the aggregate dictator allocation to their co-player in our baseline where there are no group affiliations and in the two Group treatments for each of the 3 phases. We focus first on phase 1 and use Wilcoxon ranksum tests to make comparisons across treatments and Wilcoxon signed rank tests to make comparisons within treatments. We first note that there is an in-group bias in behaviour (IGB) in both Group treatments: $CP(own)$ is significantly greater than $CP(other)$ (respectively in Group and Group-Fine, $p < 0.00001$; $p = 0.0084$) in Phase 1.

Table 2. Mean dictator allocations

| | Obs. | Recipient's group | | | | | |
|------------|------|-------------------|------------------|------------------|------------------|------------------|------------------|
| | | Phase 1 | | Phase 2 | | Phase 3 | |
| | | Own | Other | Own | Other | Own | Other |
| Baseline | 38 | 29.47 (13.74) | | 30.39 (15.74) | | 29.61 (16.54) | |
| Group | 71 | 37.75 (15.30) | 28.83 (15.02) | 40.00 (15.17) | 28.45 (15.06) | 38.87 (17.51) | 29.23 (16.64) |
| Group-Fine | 78 | 34.62 (14.07) | 29.94 (14.78) | 37.26 (14.78) | 29.69 (13.75) | 37.08 (15.30) | 28.17 (14.56) |

Figures in parentheses are standard deviations. Dictators and recipients in the Baseline do not have a group identity. All participants have an endowment of 50 tokens each. The size of the pie the dictator splits is 80 tokens in all cases.

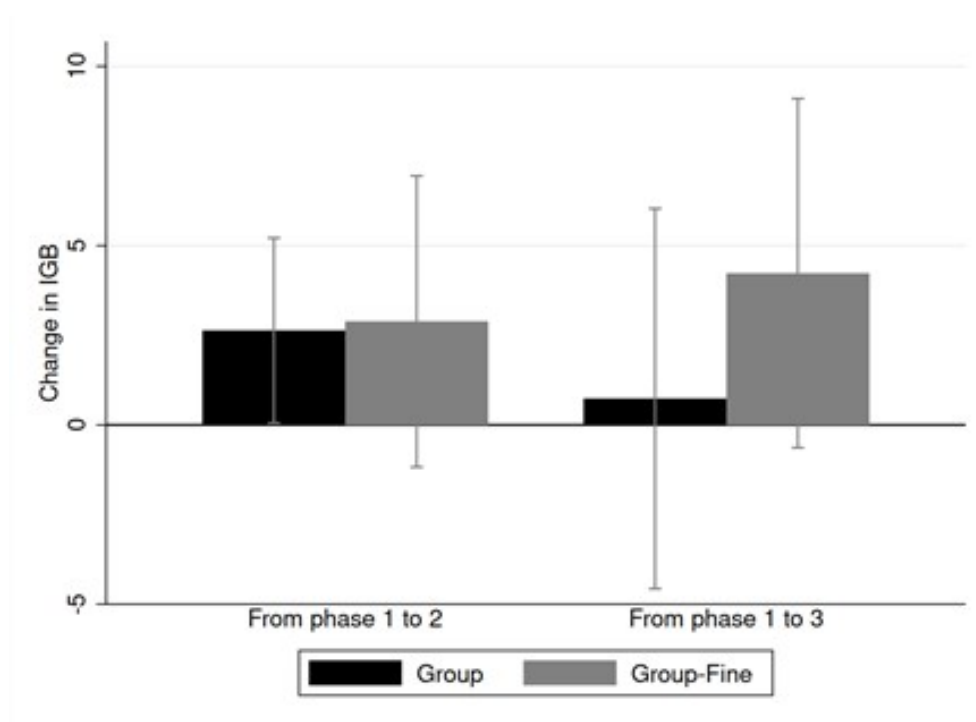
To test H1 on the social identification sources of this bias, we note that the phase 1 baseline allocation CP is very similar to $CP(other)$ in both the Group treatments. This goes against the

⁸ There was a maximum time limit of 40 minutes after which subjects who had not yet completed the experiment were automatically ejected from the study by Prolific, and no data from them were recorded.

alternative out-group hate explanation for IGB and is consistent with weak or no overall group identification in the baseline under social identification theory. Furthermore, $CP(own)$ is higher than CP as predicted by social identification as the source of the bias. $CP(own)$ in Group is significantly higher than the baseline CP ($p = 0.0098$), but while $CP(own)$ in Group-Fine is higher than the baseline CP this is not significantly higher ($p = 0.2136$). During phase 1 there is no reason to distinguish Group and Group-Fine, and when we combine $CP(own)$ in these two Group treatments, $CP(own)$ is significantly higher than the baseline CP ($p = 0.0381$).

Result 1 (support for H1, social identification and the in-group bias): Dictator allocations to own group co-players are higher than the allocation to co-players from the other group, and this is due to higher $CP(own)$ in the Group treatments than CP in the baseline.

Figure 1. Aggregate change in IGB with 95% confidence intervals



We turn now to the crowding-out/in hypotheses and begin with H2, the aggregate test. The left part of Figure 1 shows the average change in IGB between phase 1 and phase 2 in the two Group treatments, along with the 95% confidence intervals. The average change is similar in the two treatments and the difference between them is not significant (2.63 vs. 2.88, $p = 0.6782$). Result 2 follows.

Result 2 (in support of H2, no aggregate crowding out): The introduction of the fine has no aggregate effect on the in-group bias in behaviour.⁹

With respect to the individual crowding-out hypothesis H3, the top panel of Table 3 reports summary statistics of the magnitude of the IGB, $CP(own) - CP(other)$, of those who revealed such a bias in phase 1 in the Group and Group-Fine treatments and also how this group of subjects' IGB evolves in phases 2 and 3. We compare the change in $CP(own) - CP(other)$ between phase 1 and 2 in Group and Group-Fine for this group of IGB individuals in phase 1: the change in IGB in Group-Fine is significantly less than the change in IGB in Group ($p = 0.0482$ when looking at absolute changes, $p = 0.0396$ when looking at percentage changes).

Table 3. Mean change in favouritism conditional on level of favouritism in Phase 1

| | Obs. | Group favouritism | | | Change in group favouritism | | % Change in group favouritism | |
|-----------------------------------|------|-------------------|------------------|------------------|-----------------------------|------------------|-------------------------------|----------------|
| | | Phase 1 | Phase 2 | Phase 3 | From Phase 1 to 2 | 1 to 3 | From Phase 1 to 2 | 1 to 3 |
| In-group favouritism (IGB) | | | | | | | | |
| Group | 32 | 22.28 (18.01) | 24.53 (18.59) | 17.97 (31.13) | 2.25 (14.08) | -4.31 (29.84) | 0.27 (0.91) | 0.09 (1.89) |
| Group-Fine | 24 | 21.46 (17.48) | 15.38 (21.79) | 17.29 (20.27) | -6.08 (19.78) | -4.17 (23.62) | -0.22 (1.27) | -0.04 (1.2) |
| Equal allocations (EQT) | | | | | | | | |
| Group | 35 | 0 (0) | 1.86 (5.16) | 3.43 (12.59) | 1.86 (5.16) | 3.43 (12.59) | - | - |
| Group-Fine | 45 | 0 (0) | 6.24 (16.57) | 4.18 (14.65) | 6.24 (16.57) | 4.18 (14.65) | - | - |

Figures in parentheses are standard deviations. There were 4 (9) individuals in Group (Group-Fine) who displayed OGBSP in Phase 1. Given the small number of observations here, we do not conduct any analysis of the behaviour of these individuals.

Parametric individual regressions in Table 4 also support H3 and provide additional evidence in support of H2. Table 4 gives the OLS regressions for the Group and Group-Fine treatments with the change in $CP(own) - CP(other)$ between phase 1 and 2 as outcome variable. In

⁹ The analogous result in the complementary experiment where the fine is levied on IGB behaviour is the opposite: the fine has an aggregate effect because it significantly reduces IGB behaviour. We cannot conclude from this that there was no crowding-out, but if it exists, it is clearly weaker for the fine on IGB behaviour than is the fine on non-IGB behaviour.

column (1) we just consider a dummy for the Group-Fine treatment, where the Group treatment is the omitted category. Column (2) has the full set of interactions between this dummy and two dummies for subjects who revealed IGBSP and OGBSP respectively in phase 1. EQTSP subjects in phase 1 are the omitted category for these dummies. Column (3) adds socio-demographic controls at the individual level.¹⁰ Results from column (1) show that on average the fine does not have a significant impact on IGB at the individual level (i.e., further support for H2). However, in column (2) the coefficient on the interaction between the Group-Fine treatment and displaying IGB in phase 1 is negative and significant at 5% level in support of H3. This result is robust to the additional controls of column (3).

Table 4. Regression on change in in-group bias between phase 1 and 2

| | Change in favouritism | | |
|---------------------|-----------------------|---------------------|---------------------|
| | (1) | (2) | (3) |
| Group-Fine | 0.251 (2.449) | 4.387 (3.308) | 4.45 (3.704) |
| IGBSP | | 0.393 (3.589) | -1.999 (4.052) |
| Group-Fine × IGBSP | | -12.72** (5.162) | -13.56** (5.792) |
| OGBSP | | 10.64 (7.746) | 12.87 (8.532) |
| Group-Fine × OGBSP | | -6.887 (9.419) | -9.216 (10.44) |
| Constant | 2.634 (1.808) | 1.857 (2.481) | -7.962 (16.21) |
| Individual controls | NO | NO | YES |
| Obs. | 149 | 149 | 149 |

Standard errors in parentheses. *** p < 0.01, ** p < 0.05, * p < 0.10

¹⁰ The controls are gender, age, education, employment status, political and economic opinions and a measure of performance in the previous real effort task. None of these control variables is statistically significant.

Result 3 (in support of H3, individual crowding out): For players who revealed IGBSP in phase 1, IGBSP is crowded-out in Group-Fine in phase 2 as compared with Group.¹¹

We turn to whether the crowding-out that we observe can be attributed to a fall in social identification with one's own group (H4). Table 5 presents the average $CP(own)$ and $CP(other)$ for subjects who reveal IGBSP in phase 1. In Group-Fine, IGB decreases from Phase 1 to Phase 2 due to a decrease in $CP(own)$ and an increase in $CP(other)$. However, neither change is significant when compared with the change in the baseline ($CP(own)$: $p = 0.2329$; $CP(other)$: $p = 0.1890$).

Table 5. Mean dictator allocations for subjects who reveal IGB in phase 1

| | Obs. | Phase 1 | | Phase 2 | | Phase 3 | |
|------------|------|------------------|------------------|------------------|------------------|------------------|------------------|
| | | Own | Other | Own | Other | Own | Other |
| BASELINE | 38 | 29.47 (13.74) | | 30.39 (15.74) | | 29.61 (16.54) | |
| Group | 32 | 43.91 (16.3) | 21.62 (12.22) | 46.41 (15.77) | 21.88 (13.84) | 43.59 (20.21) | 25.62 (19.12) |
| Group-Fine | 24 | 38.12 (15.24) | 16.67 (10.39) | 36.54 (16.47) | 21.17 (13.26) | 35.21 (16.97) | 17.92 (12.76) |

Figures in parentheses are standard deviations.

Table 6 gives individual OLS regressions on H4. Columns (1) and (2) present regressions of $CP(own)$ in Phase 2 with $CP(own)$ in Phase 1, treatment dummies (excluded treatment: BASELINE), dummies for IGB and OGB in Phase 1 (excluded category: EQB) and their interactions with Group-Fine as explanatory variables. Columns (3) and (4) present the corresponding regressions for $CP(other)$.

¹¹ As noted in footnote 5, the test for crowding-out in the complementary experiment is weaker than in the main experiment because EQTSP individuals cannot reveal adjustments downwards in their EQTSP. Nevertheless, there is no evidence of individual crowding out because the size of the bias does not shrink among EQTSP and OGBSP compared with Group (test for the joint $p = 0.3799$). Together with the analogous result in the complementary experiment to Result 2, reported in footnote 11, this suggests crowding-out was weak at best in the case of fines designed to discourage IGB. This cannot be because EQTSP is more salient as the 'correct' social preference in this experiment because the numbers with IGBSP and EQTSP are about the same. It is possible though that the fine is associated with a market-type intervention (i.e. it puts a price on a particular kind of behaviour) and markets are known from other experiments to encourage equal treatment (see Hargreaves Heap *et al.*, 2013). In this way, the fine may actually reinforce the equal treatment behaviour even though it sets up the conditions where intrinsic is no longer necessary to explain behaviour and so might in other circumstances diminish.

Table 6. Regressions on $CP(own)$ and $CP(other)$ in phase 2

| | (1) Own | (2) Own | (3) Other | (4) Other |
|--------------------------------------|---------------------|---------------------|---------------------|---------------------|
| Allocation to own group in Phase 1 | 0.723*** (0.057) | 0.709*** (0.062) | | |
| Allocation to other group in Phase 1 | | | 0.740*** (0.059) | 0.717*** (0.063) |
| Group | 11.59* (6.204) | 10.79 (6.64) | -2.134 (5.829) | -3.649 (6.188) |
| Group-Fine | 15.34** (7.135) | 15.44** (7.598) | -2.817 (6.707) | -3.564 (7.084) |
| IGBSP | 4.173 (2.672) | 3.632 (2.879) | -2.209 (2.526) | -1.282 (2.728) |
| Group-Fine \times IGBSP | -9.436** (3.763) | -10.13** (4.029) | 3.642 (3.54) | 3.382 (3.771) |
| OGBSP | 6.012 (5.599) | 5.202 (5.863) | 0.58 (5.394) | -1.337 (5.623) |
| Group-Fine \times OGBSP | -9.407 (6.856) | -7.978 (7.233) | -3.15 (6.487) | -0.919 (6.813) |
| Constant | -1.096 (6.971) | 9.235 (13.57) | 10.22 (6.596) | 29.12** (12.69) |
| Individual Controls | NO | YES | NO | YES |
| Obs. | 187 | 187 | 187 | 187 |

Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

The interaction between the Group-Fine treatment and IGB in phase 1 on allocations to co-players from one's own group is negative and significant at 5% (column 1), while the same interaction is not significant, although positive, when measured on allocations to co-players from the other group (column 3). This is robust to adding individual characteristics as controls (columns 2 and 4).

Result 4 (in support of H4): There is evidence from individual-level regressions that crowding-out of IGBSP in phase 2 in Group-Fine is driven by a reduction in $CP(own)$ more than by an increase in $CP(other)$.

To test H5 (the possibility of crowding-in of IGBSP), we consider the 20 out of 45 individuals who reveal EQTSP in phase 1 and also adjust to the fine in phase 2 by moving to IGB. On average, these subjects move to an IGB of 17.30 in phase 2 in Group-Fine. This is not a marginal change to avoid the fine: we can reject the hypothesis that $IGB=1$ (i.e., a marginal adjustment) in phase 2 ($p < 0.00001$). Furthermore, those who reveal IGBSP in phase 1 in the Group treatments have an average IGB of 21.92 and the difference between this and 17.30 is only weakly significant ($p = 0.0916$).

Result 5 (supporting H5): Those who reveal EQTSP in phase 1 and go on to reveal IGB in phase 2 adjust non-marginally to the fine in Group-Fine.

Finally, we consider how many of these results persist in phase 3. Result 1 translates completely, as sign rank tests reveal that in the Group treatments $CP(own)$ is still significantly greater in phase 3 than $CP(other)$ (for Group: $p = 0.0002$, for Group-Fine: $p < 0.0001$) and ranksum tests relative to the baseline on $CP(own)$ provide evidence that it is still because of social identification rather than out-group hate (for Group: $p = 0.0088$, for Group-Fine: $p = 0.0323$).

Turning to the impact of the fine and our hypotheses on crowding out/in, we now consider aggregate changes between phase 1 and 3. At the aggregate level, the fine still has no significant impact on IGB (see Figure 1: 0.73 vs. 4.23, $p = 0.4301$). To assess Result 3, we examine if the change in IGB for subjects who reveal IGBSP in phase 1 in Group-Fine is statistically significant by comparing it with the equivalent change in the Group treatment (see Table 3). It turns out that individual crowding-out does not persist, both when looking at absolute changes ($p = 0.9498$) and at percentage changes ($p = 0.9230$). Regarding Result 5, we turn to subjects in Group-Fine who revealed EQTSP in phase 1 and IGB in phase 2. In phase 3 they display an average in-group bias of 8.90. A sign rank test on IGB for these subjects between phase 1 and 3 reveals that the difference is significantly positive ($p = 0.0078$). Therefore, we can conclude that Result 5 holds also in phase 3: i.e., the crowding-in of IGBSP still persists among those who initially revealed EQTSP in phase 1.

4. Discussion and conclusion

In our Group treatments, we find that subjects exhibit an IGB on average in phase 1. They give more to someone from their own group than to someone from another group. Since there is no

material incentive to treat other people differently depending on which group they belong to, this bias reveals IGBSP. This finding is consistent with that of many experiments where IGB has been found. We also find that this bias might be explained by social identification theory. The bias arises in the Group treatments because the allocation to someone from own group rises relative to the allocation in the BASELINE where there are no group affiliations (and the allocation to someone from the other group is no different to the BASELINE allocation). As would be expected from social identification theory when individuals identify more strongly with members of their own group than those from other groups, subjects treat own group members especially kindly compared with how other people are generally treated. Since social identification theory has been found helpful in explaining other behaviours in economics, this finding too is broadly consistent with the literature (e.g., see Akerlof and Kranton, 2005). In these respects, our experiment coheres with what is known from other studies.

Our contribution is the test for the crowding-out of IGBSP when material incentives towards IGB are introduced. This, to our knowledge has not hitherto been considered or tested. We are the first to examine this possibility. IGBSPs are crowded-out in the aggregate. At the individual level, the picture is more complicated. Those who initially reveal IGBSP in phase 1, exhibit less IGB after the fine is introduced: their IGBSPs are crowded-out by the introduction of the fine. However, there is some evidence that individuals who did not have IGBSP initially and adjusted to the fine develop IGBSP after the fine: i.e., IGBSPs are crowded-in for this set of individuals. Further there is some evidence that this crowding-in persists into phase 3. Nevertheless, although there are these heterogeneous effects on social preferences, the balance both in phase 2 and phase 3 favours the crowding-out effect in the aggregate because the crowding-out of the subjects who revealed IGBSP in phase 1 is sufficiently large and the number of subjects who are willing to pay the fine is sufficiently high (32 out of 54)¹² to compensate the crowding-in of those who adjust to the fine. Furthermore, the crowding-out that we observe is consistent with a weakening of the social identification origins of this bias. It occurs because the special generosity shown to own group members shrinks in phase 2.

These are important results in two respects.

First, they caution against the use of material incentives to encourage the in-group bias because it produces an offsetting crowding-out of the intrinsic motivation (IGBSP) towards such behaviour. In our experiment, this crowding-out is such that there is no effect in the aggregate

¹² Out of the 45 (9) subjects who revealed EQTSP (OGBSP) in phase 1, 25 (7) were willing to pay the fine in phase 2, i.e., they displayed EQTSP or OGBSP in phase 2.

from the introduction of the material encouragement towards IGB. The policy is ineffective. This has special relevance to nationalist policies and yields a paradox where such policies undermine their own foundations, but it has relevance for any policy that seeks to influence the in-group bias with a change in material incentives.

Second, we find evidence in support of social identification theory in our experiment: both in the explanation of the IGB and in the crowding-out of social identification through the introduction of material incentives. The latter is important because it suggests that the social preferences that arise from social identification are not always stable. In short, such preferences are not the bedrock that economists sometimes assume preferences are (or ought to be).

Acknowledgements

The authors thank participants at the ESA Global Online Meetings in 2020 for helpful comments and suggestions. Funding from Appalachian State University and King's College London is gratefully acknowledged. The study was exempted from review by the IRB at Appalachian State University: Study # 20-0260.

References

- Akerlof, George A., and Rachel E. Kranton (2000) "Economics and Identity", *Quarterly Journal of Economics*, 115(3), 715-753.
- Akerlof, George, A., and Rachel E. Kranton (2005) "Identity and the Economics of Organizations", *Journal of Economic Perspectives*, 19 (1), 9-32.
- Bowles, Samuel, and Sandra Polania-Reyes (2012) "Economic Incentives and Social Preferences: Substitutes or Complements?", *Journal of Economic Literature*, 50(2), 368-425.
- Chen, Yan, and Sherry Xin Li (2009) "Group Identity and Social Preferences", *American Economic Review*, 99(1), 431-57.
- Chen, Daniel L., Martin Schonger, and Chris Wickens (2016) "oTree – An open-source platform for laboratory, online, and field experiments", *Journal of Behavioral and Experimental Finance*, 9, 88-97.
- Deci, Edward L. (1975) *Intrinsic Motivation*, Plenum Press: New York.
- Erkal, Nisvan, Lata Gangadharan, and Nikos Nikiforakis (2011) "Relative Earnings and Giving in a Real-Effort Experiment", *American Economic Review*, 101(7), 3330-48.

- Frey, Bruno (1997) *Not Just for the Money: an economic theory of personal motivation*. Cheltenham: Edward Elgar.
- Gneezy, Uri, and Aldo Rustichini (2000) “A Fine is a Price”, *Journal of Legal Studies*, 29(1), 1-17.
- Hargreaves Heap, Shaun P., and Daniel John Zizzo (2009) “The Value of Groups”, *American Economic Review*, 99(1), 295-323.
- Hargreaves Heap, Shaun P., Jonathan HW Tan, and Daniel John Zizzo (2013). “Trust, inequality and the market”, *Theory and Decision*, 74(3), 311-333.
- Lucas, Jr., Robert E. (1976) “Econometric Policy Evaluation: A Critique”, In Bruner, Karl and Alan Meltzer (eds.) *The Phillips Curve and Labor Markets*. Carnegie-Rochester Conference Series on Public Policy. 1. New York: American Elsevier, 19–46.
- Shayo, Moses (2020) “Social Identity and Economic Policy”, *Annual Review of Economics*, 12, forthcoming.
- Stigler, George J., and Gary S. Becker (1977) “De Gustibus Non Est Disputandum”, *American Economic Review*, 67(2), 76-90.
- Tajfel, Henri, and John Turner (1979) “An integrative theory of intergroup conflict”, In Worchel, Stephen and William Austin (eds.) *The social psychology of intergroup relations*, Brooks/Cole Pub. Co., 33–47.