European Economic Review 90 (2016) 375–402

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

Discrimination in the laboratory: A meta-analysis of economics experiments

Tom Lane

Department of Economics, University Park, University of Nottingham, NG7 2 RD, United Kingdom

ARTICLE INFO

Available online 29 December 2015

JEL classifications:
C92–Laboratory Experiments (Group Behaviour)
D03 – Behavioural Microeconomics: Underlying Principles

Keywords:
Discrimination
Meta-analysis

ABSTRACT

Economists are increasingly using experiments to study and measure discrimination between groups. In a meta-analysis containing 441 results from 77 studies, we find groups significantly discriminate against each other in roughly a third of cases. Discrimination varies depending upon the type of group identity being studied: it is stronger when identity is artificially induced in the laboratory than when the subject pool is divided by ethnicity or nationality, and higher still when participants are split into socially or geographically distinct groups. In gender discrimination experiments, there is significant favouritism towards the opposite gender. There is evidence for both taste-based and statistical discrimination; tastes drive the general pattern of discrimination against out-groups, but statistical beliefs are found to affect discrimination in specific instances. Relative to all other decision-making contexts, discrimination is much stronger when participants are asked to allocate payoffs between passive in-group and out-group members. Students and non-students appear to discriminate equally. We discuss possible interpretations and implications of our findings.

1. Introduction

Meta-analysis – a commonplace technique in medical science, psychology and, to a growing extent, economics – holds advantages over literature review in terms of objectivity and analytical rigour. In recent years, the experimental economics literature appears to have reached a critical mass at which researchers are finding meta-analyses useful.[1] The benefit of these works is that, by aggregating data across a large number of experiments and exploiting natural between-study design variation, they pinpoint behavioural regularities and the variables that modify them more precisely than could be done through qualitative review.

We run a meta-analysis on the body of studies investigating discrimination in lab and lab-in-the-field experiments, a sub-literature which has certainly reached the necessary critical mass for such a venture. Economists' interest in discrimination has been strong ever since Becker (2010), and with the growth of experimental economics in the last two decades, experiments have emerged as a popular complement to survey-based econometric studies.

E-mail address: lextl9@nottingham.ac.uk

[1] Several meta-analyses of economics experiments have been released in recent years, including: Engel (2007) – oligopoly experiments; Prante et al. (2007) – Coasean bargaining; Jones (2008) – group cooperation in prisoners' dilemmas; Hopfensitz (2009) – the effects of reference dependence and the gambler's fallacy on investment; Percoco and Nijkamp (2009) – time discounting; Weizsäcker (2010) – social learning; Engel (2011) – dictator games; Johnson and Mislin (2011) – trust games.

http://dx.doi.org/10.1016/j.euroecorev.2015.11.011
0014-2921/

These experiments create a controlled environment and therefore allow much cleaner measurements of discrimination than the analysis of naturally-occurring data, avoiding such problems as omitted variable bias and reverse causality. Furthermore, by testing for a very fundamental and general form of discrimination – simply, whether subjects treat others differently depending on which group those others belong to – experimental economists can produce findings of interest not only to their own discipline but also across the social sciences. Also, through the use of incentives, experiments hold a key advantage over questionnaire-based measures of discrimination, in that they elicit revealed rather than reported discrimination.

Psychologists had already been studying discrimination in the lab for decades, and experimental economists have drawn on their knowledge, particularly regarding the minimal group paradigm. This technique was first introduced by Tajfel et al. (1971) and has spawned a huge body of experiments wherein group identity is artificially induced in the laboratory. This is often done by, in a preliminary phase of an experiment, asking subjects to state their preference for one artist over another, or to randomly draw a colour. The experimenter then splits the subject pool into groups according to their art preference, or the colour they have drawn, and makes it known to participants that the division is based on these differences. Subsequent stages of such experiments involve interaction tasks between the groups and find discrimination surprisingly (at least to the early researchers) often.

To study discrimination, experimental economists set up games such as the dictator game, the trust game or the prisoner's dilemma, and invite a subject pool segregated along the lines of a particular identity-based characteristic (or else generate this segregation with artificial groups). They make subjects aware of the group affiliation of those they interact with, and then measure how their behaviour varies according to whether individuals they are interacting with share their identity (are in-group) or do not (are out-group).

The number of economics experiments of this type has grown rapidly since the turn of the century and now encompasses substantial diversity across several dimensions. Even after omitting many papers which investigate discrimination but do not meet our inclusion criteria devised to ensure a consistent approach (see Section 2), we are left with a dataset consisting of 441 experimental results (significant and null) from 77 studies – more data than most of the other experimental economics meta-analyses have had. In order to aid the progression of this literature, it is worth taking stock of what has been found to date, particularly as casual inspection reveals non-uniformity in the results; the strength of discrimination found against out-groups varies considerably, and some experiments even find discrimination in the opposite direction, i.e. against the in-group.

The aim of this meta-analysis is both to yield broad insights on discrimination and to inform the designers of future experiments testing for it. We first investigate the average strength of discrimination across the literature. We then inquire how it tends to vary according to specific experimental characteristics.

In particular, we are interested in whether the strength of discrimination depends on the type of identity being investigated. Comparing the level of discrimination between artificial (i.e. minimal) groups and various types of natural groups (such as those based on ethnicity, nationality, religion, gender and social/geographical affiliation) is particularly interesting. One might expect 'minimal' groups to yield minimal levels of discrimination. However, it is also conceivable that artificial identity inducement confers an experimenter demand effect in favour of discrimination, or that the experimental priming of sensitive natural identities reduces subjects' desire to discriminate owing to a preference not to engage in socially unacceptable behaviour. Evidence for these possibilities, in the form of relatively strong discrimination in artificial group experiments, could have implications for the external validity of certain experiments.

A further interesting question is whether the strength of discrimination varies according to the type of decision subjects are asked to make. This has implications in terms of the real-world circumstances in which discrimination can be most expected to appear and for the generalisability of findings.

We further ask whether experiments with students reveal greater or lesser discrimination than those with non-students. This is also important for the external validity of findings, and is a question worth pursuing as some studies (e.g. Bellemare and Kröger, 2007; Anderson et al., 2013) have found students are not entirely representative of wider populations in economics experiments.

This meta-analysis also aims to shed light on the motivations behind discrimination. Some experiments have been designed specifically to distinguish between taste-based discrimination and statistical discrimination – the two models that continue to dominate the theoretical literature in economics. The taste-based model, proposed by Becker (2010), entails individuals gaining direct utility from the act of discriminating against out-groups. Meanwhile, according to theories of statistical discrimination – beginning with Arrow (1972) – individuals aim to maximise their own payoffs given their beliefs and expectations about others' characteristics and behaviour, and discrimination occurs when those beliefs and expectations vary depending on the group to which the others belong. Understanding the relative importance of these two motivations will improve the focus of future research and the design of policies aimed at combating discrimination.

Finally, we include a subsection on experiments investigating gender discrimination. Gender is unique amongst the identity types in having the same two groups in each experiment. It is therefore simple to make a clean comparison between male-to-female discrimination and female-to-male discrimination.

In summary, the meta-analysis presented below aims to address the following questions: (1) What is the general pattern of discrimination across the literature? (2) How does the level of discrimination vary according to the type of identity groups are based upon? (3) How does the level of discrimination depend upon the decision-making context? (4) Do students discriminate any more or less than non-students? (5) Does the experimental literature provide more support for taste-based or statistical theories of discrimination? (6) In gender experiments, how does male-to-female discrimination compare with female-to-male discrimination?

Our main results, presented in Section 3, are as follows. (1) We find a moderate tendency towards discrimination against the out-group, with a majority of null results across the literature. (2) The strength of discrimination against the out-group does vary according to the type of group identity subjects are divided by. It is greater when identity is artificially instilled in a subject pool than when it is divided by nationality or ethnicity – minimal groups, it seems, are not so minimal after all. Discrimination is even stronger, though, when participants are divided into socially or geographically distinct groups. (3) The extent of discrimination against the out-group also depends on the role participants are given in an experiment: when subjects are asked to allocate payoffs between inactive players belonging to the in-group and out-group, it is stronger than in any other decision-making context. (4) Students do not appear to be differently inclined towards discrimination than non-students. (5) We find evidence in support of both taste-based and statistical discrimination. Tastes appear to drive the general tendency for discrimination against the out-group, but individual studies have found beliefs to affect discrimination. (6) In gender discrimination experiments the tendency for discrimination against the out-group is reversed, as subjects demonstrate slight but significant favouritism towards the opposite gender. Discriminatory behaviour in these experiments does not differ significantly between males and females. We discuss possible interpretations of these results in depth in Section 4.

We are aware of only one other meta-study attempting to analyse the experimental discrimination literature – Balliet et al. (2014),[2] who take 214 estimates of discrimination from 78 studies. There is little overlap between our samples; Balliet et al. take studies from across the social sciences but their search and inclusion criteria result in most of the experimental economics literature on discrimination not being included (26 of our studies – around a third – feature in Balliet et al.'s sample). They exclude decision-making contexts which we consider, such as being the second mover in a sequential game or a third-party allocator. They also exclude interactions between gender groups.

The present study and that of Balliet et al. can be viewed as complements. Through focusing only on economic experiments, we enhance comparability and eliminate some studies using methodological elements that may not be acceptable to some social scientists. Our focus on the economic theories of taste-based and statistical discrimination differentiates our study from Balliet et al., who investigate psychological theories of discrimination. Throughout our analysis we compare our results to theirs. Their paper finds a similar overall tendency for discrimination to what we do. They find the extent of discrimination not to differ significantly between settings of natural and artificial identity, but do not split natural identity into subcategories as we do. The clearest difference in results between the two studies is that Balliet et al. find discrimination is stronger by decision-makers who move simultaneously than by first movers in sequential exchanges, while we do not find it significantly differs between these settings.

2. Methodology and criteria for inclusion

We chose to restrict our study to the experimental economics literature. Almost all of the economics experiments have been conducted in the last 15 years and can reasonably be expected to have followed comparable procedures, which is important in a meta-analysis. We define an economics paper as follows: it must either have been published in an economics journal or have as at least one of its authors a person trained in economics or a business-related discipline, or who has at least once held a position in an economics or business-related department. Furthermore, we exclude economics papers which, it is clear to the reader, exhibit a breach of standard experimental economics practice – most notably, deception. For inclusion, an experiment must involve interaction between individuals whose decisions determine real material payoffs for participating players. In other words, it must be incentivised.

A serious pitfall meta-analyses can face is publication bias, also named the 'file drawer problem'. Because null results are less likely to be published than significant ones, a meta-analysis risks including a disproportionately low number of studies finding small or no effects (Rosenthal, 1979; Rothstein et al. 2006). This can lead to an overestimation of average effect sizes. It can also, if null results are particularly unlikely to be published when combined with certain other features of a study, result in the meta-analysis overestimating the relationship between strong effects and these features; in our case, for instance, if null results in trust games were never published but null results in other games sometimes were, we would be in danger of estimating a spuriously strong relationship between trust games and significant results. To minimise such bias, a good meta-analysis should conduct the most thorough literature search possible in order to find all applicable studies,

[2] Although nothing approaching a full meta-analysis of the in-group-out-group literature had previously been conducted, several social psychology meta-studies have investigated specific phenomena within it. Saucier et al. (2005) analysed research measuring the degrees to which subjects would help white and black people; while not finding statistically significant aggregate discrimination against black people, they showed it increased in emergency situations and cases where helping was more difficult or risky. Bettencourt et al. (2001) found high-status groups exhibited more in-group bias than low-status groups. Fischer (2010) concluded discrimination in minimal group experiments was stronger in countries whose societies are considered more individualistic. Aberson et al. (2000) found greater in-group bias amongst individuals with higher self-esteem. Robbins and Krueger (2005) found social projection, 'the tendency to expect similarities between oneself and others', to be stronger towards in-groups than out-groups, and that this effect was amplified with artificial groups relative to natural ones. Although interesting, many of the studies included in these meta-analyses are considerably different from those we consider – often they do not relate specifically to economic behaviour, and even if they do they may not be incentivised.

whether published or not. Our approach was threefold. In late 2013, we conducted RePEc searches for the keywords, 'Discrimination experiment', 'Identity experiment', 'Ingroup experiment' and 'Outgroup experiment', and carefully sifted through the output for candidate studies. We then followed the references and citations of all papers identified as relevant. Finally, we checked our list of included studies against that of Balliet et al. (2014); this step added one study (Spiegelman, 2012).[3] One feature of the literature we meta-analyse is that studies tend to include various different treatments, and therefore report multiple results. This may act as a further curb on publication bias – insignificant findings make their way into papers alongside more interesting significant results (indeed, it turns out the majority of results in our dataset are null).[4]

Previous meta-analyses in experimental economics such as Engel (2011) and Johnson and Mislin (2011), which focus on a single game type, are able to use the average behaviour of subjects (amount sent in the dictator or trust game) as a continuous dependent variable, with one observation and an associated standard error for each treatment. In our case, we are pooling across different game types and therefore need a way of transforming the data to make meaningful comparisons between these settings. Our variable of interest is the difference between decision-makers' behaviour towards their in-group and their out-group, whilst all other aspects of the experimental design are held constant – in essence, the discrimination effect size. There is typically one observation per every two treatments (one in-group and one out-group treatment) for each type of player active in the given game. The exception is when a decision-maker interacts with both the in-group and the out-group in the same treatment (either by making one decision which simultaneously affects both, or by playing in the same role twice), in which case a within-treatment measure of discrimination is available.[5] The ideal approach would be to record an effect size for each comparison, and we attempt to do this. Consistent with Balliet et al. (2014), the measure we use is Hedges' unbiased d: the mean difference in behaviour towards the in-group and the out-group, divided by the pooled standard deviation, with a minor correction for sample size (Hedges and Olkin, 1985).

However, a substantial number of studies do not report sufficient data for us to calculate effect sizes. This is particularly the case with null results, as when a difference is not significant authors are less likely to report the test statistic from which an effect size could be derived. We sent data requests to the authors of all papers for which we could not construct the measure using information provided in the paper. After receiving data from 22 of the 36 sets of contacted authors, we ended up with effect sizes on 364 of our 441 data-points. We therefore also employ a binary dependent variable, recording simply whether, for each comparison, behaviour significantly favours the in-group over the out-group at the 5% level.[6] The effect size is the inferior dependent variable in that it restricts the sample and may lead to greater under-representation of null results; but the superior one in terms of information content.

For simplicity, we define 'discrimination' as discrimination against the out-group, and 'out-group favouritism' as discrimination against the in-group, and will use these terms hereafter. Unlike some, we make no distinction between nepotism and discrimination; any result of favouritism towards one group relative to a second can equivalently be interpreted as discrimination against the second group. We therefore conceptualise 'discrimination' (against the out-group) as something which can be measured on a continuum with positive and negative values. When discussing average effect sizes, we will describe a relatively low value as indicating 'lower' or 'weaker' discrimination, even if it is driven by highly negative effect sizes (i.e. even if it is driven by instances of strong discrimination against the in-group).

For an observation to meet our inclusion criteria, there must be an in-group and out-group, clearly defined on the basis of categorisation by a discrete identity-relevant variable, such as ethnicity, gender or – as with artificial groups – the preference for a particular artist or the colour randomly drawn. There must be controlled interaction within and between the groups, and decision-makers must be aware that they are interacting with individuals belonging to their in-group or out-group. We only consider an in-group to be appropriately defined as such if every one of its members shares the same categorisation as the decision-maker on the basis of the relevant variable. For an out-group to be appropriately so-defined, every member must take a different categorisation from the decision-maker. It is not required that all members of an out-group take the same categorisation as each other. For instance, Guillen and Ji (2011) use as their two groups Australian and non-Australian. In this case, for an Australian decision-maker the Australians are the in-group and the non-Australians the out-group, but for a non-Australian the other non-Australians should not count as their in-group. We then only record the observed behaviour of the appropriately defined group, the Australians in this example. Occasionally, we are forced to make a subjective decision on what can reasonably be considered a group. For example, from Chen et al. (2011), which splits its US-based sample into white and Asian students, we record the behaviour of the white 'group' but not that of the Asians, as we believe that in American society white people can appropriately be defined as comprising a shared ethnicity, whilst

[3] The Balliet et al. project was not in the public domain when we embarked upon ours, and we were unaware of it. We designed our search and inclusion criteria independently of theirs. However, learning of their meta-analysis provided the perfect opportunity to test the thoroughness of our search for studies. That Balliet et al. include only one study which fits our inclusion criteria but which we had not independently found suggests it is unlikely we have missed many applicable papers.

[4] The number of observations generated by a single paper varies from 1 to 24, with Chen et al. (2014) providing the most.

[5] For a game to meaningfully measure discrimination, and therefore for us to include it, it must be possible to unambiguously rank the decision-maker's available actions in terms of how favourable they are to the decision-maker's partner. Certain coordination games cannot be included, since whether one action is more favourable depends upon the move a partner simultaneously makes. In Appendix A, Table A.2 we list all the game types included in our sample, and explain how they measure discrimination.

[6] We also do this for out-group favouritism, recording whether or not behaviour significantly favours the out-group over the in-group at the 5% level, and run separate regressions on this. These are reported in Appendix C, Table C.1.

those of Asian descent comprise a mixture of ethnicities.[7] Papers such as Falk and Zehnder (2007) which do not have clear groups but measure each subject's position on a scale of social distance, based on a continuous variable, are not included.

If an experimental design splits the sample up into more than two separate groups, on the basis of a single identity-relevant variable, we record separately how each group treats each other group relative to its own. If such a paper reports that Group A does not significantly discriminate against Group B or Group C but does significantly discriminate against Groups B and C combined, we record two results of no discrimination rather than one result of discrimination; and in the main text of this paper we report our results using this approach. We do this because, although Groups B and C combined could represent a single out-group as defined above, the experiment was set up to treat them as separate out-groups. Similarly, we do not include the reported results of statistical tests run on data pooling two or more treatment pairs. These are grey areas but we have re-run our main regression results for the binary dependent variables in the case of treating every result reported in our sample as an observation: this adds 16 extra data-points and does not qualitatively change our findings.

Sufficient data must be reported for it to be clear whether there is significant discrimination in each pair of treatments (or, when applicable, single treatment); if we cannot work out whether there is discrimination in one or more treatment pair, the whole paper is omitted from the study. This is because papers are less likely to report the results of statistical tests finding no discrimination, and if we failed to include a given study's non-results our analysis would overestimate the likelihood of this particular design finding discrimination. For similar reasons, if an experiment employs a cross-cutting design, dividing its subject pool by multiple identity types, it must report whether there is discrimination on the basis of each category. For example, an experiment which segregates the subjects by both gender and ethnicity must report, for each applicable treatment pair, whether each ethnic group discriminates against each other ethnic group or not, and also whether each gender discriminates against the other or not. Otherwise, we omit the study.

Experimenters using artificial groups generally conduct tests on pooled data; rather than reporting whether Group A discriminates against Group B and vice versa, they report whether individuals across the sample pool discriminate against out-group members. This makes sense because there is no obvious reason to doubt the relationship between two artificial groups is completely symmetrical. As such, we use pooled discrimination observations for artificial group experiments. Using similar reasoning, we also admit pooled discrimination observations for experiments dividing subjects by their real-world social groups. The pooling of certain types of data might lead to an increased chance of finding discrimination in certain experiments, which is one reason why we use the size of the sample from which the result is derived as a control variable in our regression analysis.

We limit our analysis to lab and lab-in-the-field experiments; we do not include pure field experiments, in which subjects do not know they are participants in a study. We therefore do not include the large body of field experiments in which applications are sent to employers, landlords or others to test for discrimination in markets (correspondence studies).

2.1. Analytical methods

Listed in the next subsection are descriptions of the independent variables we include in our regressions. Our basic model contains role and identity type dummies, and some controls. Because our samples are not large and most variables are dummies, we regard linear probability models (LPMs) with errors corrected for heteroskedasticity as the best specifications when employing the binary dependent variables. However, we also run as robustness checks logit models, which we report in Appendix C, Table C.2. In some cases the logits drop observations, which is a major disadvantage. Their results, however, are qualitatively similar to the LPMs. When using binary dependent variables, we treat each study within the meta-analysis as providing a cluster of observations.

When analysing the continuous dependent variable, we first use standard random effects meta-analysis procedures to determine average effect sizes for our full sample and for the subsamples based on identity type. These are simply aggregate estimates for the level of discrimination in the relevant subsample; they do not control for independent variables. The procedure takes into account that each observation has an associated standard error. It weights each observation by the inverse of this standard error, thus attaching more importance to results from larger samples and with smaller standard deviations. It then follows an unweighting process, the extent of which depends upon the heterogeneity in effect sizes. The more heterogeneity there is across effect sizes, the more equal will be the weights attached to observations with small or large standard errors (Harbord and Higgins., 2008).[8]

We then apply random effects meta-regressions, which allow the inclusion of independent variables in the analysis. These models follow the same processes of weighting and unweighting observations as described in the previous paragraph, but are otherwise standard linear regressions. Whereas with the binary dependent variable we must approach discrimination and out-group favouritism separately, the meta-regression analyses both simultaneously, since the effect sizes can be positive or negative. This can be one reason why the results of the meta-regressions may differ from those of the

[7] There were four cases where we made such subjective decisions, all listed in Appendix A. Our main results still hold regardless of the decisions we come to in these cases.

[8] The random effects approach is more suitable for our purposes than the fixed effects alternative, which excludes the unweighting step; the fixed effects process assumes there to be one true effect size across all studies, while random effects allow it to vary – the latter seems more plausible in our case, as we do not assume discrimination to be a universal constant.

linear probability regressions. Another can be the reduction in sample – therefore, when the results of the meta-regressions do not match those of the LPM regressions on discrimination, we present the LPMs re-run on the reduced effect-size sample, in order to determine whether the disparity is due to the change in sample or the change in analytical approach.

2.2. Independent variables

2.2.1. Role type dummies

We include role type dummy variables to pursue the question of how different decision-making contexts affect the extent of discrimination. The games used in this literature feature either multilateral or unilateral decision-making. When decision-making is multilateral, the outcome of the game is determined by more than one player's actions. From such situations, we identify three different role types: *First Mover* (140 observations), where one's move does not finish the game; *Second Mover* (119 observations), where one determines the final payoffs in response to the co-player(s)' actions; and *Simultaneous Mover* (66 observations), where one makes the last move of the game at the same time as one's co-player(s).

When decision-making is unilateral, the final outcome of the game is determined by one player. From these situations, we identify a further three role types. First, there is the *Dictator* (67 observations), who allocates payoffs between another player and his- or herself. Next we have third-party allocators (*Allocator*, 30). These are players who must divide a pie between two or more passive players (who, in these experiments, are members of different groups), but whose own payoff does not depend on this decision. Finally, there are players tasked with selecting a partner (from a choice of in-group and out-group participants) with whom to play a subsequent game. We label this role *Partner Chooser* (19 observations).[9]

2.2.2. Identity type dummies

A second set of dummy variables records which type of group identity a given experimental sample has been divided according to. We consider identity to have been artificially induced if researchers split subjects into groups that, prior to the experiment, did not exist – in the sense of group members sharing characteristics that are not also shared by members of other groups in the study – and the subjects are aware they have been split into these groups.[10] 49 studies in the meta-analysis investigate natural identity, 32 artificially generate it, while the remaining four contain both natural and artificial treatments. We have 272 observations for natural identity types and 169 for artificial. We subdivide the natural observations into six specific categories of natural identity.

First, we have 82 observations from 13 studies in which subjects are divided by *Nationality*. Next, nine studies investigate *Ethnicity*-based identity, adding 63 observations. A further seven studies generate 32 observations on *Gender* identity. 21 more observations are provided by five studies in which the subjects are split by *Religion*. 13 studies use a rather different approach, dividing the subject pool into groups based on real-world social and/or geographical identity. This is done in a variety of ways: for instance, using villages (Dugar and Shahriar, 2010), colleges within universities (Banuri et al., 2012) or friendship groups (Brandts and Sola, 2010). However, all such designs share the common feature that each decision-maker has a clearly distinct social and/or geographical in-group – group identity here is defined with reference to the relative frequency with which one interacts with in- and out-group members in ordinary life. The 57 observations generated by these experiments are coded under the variable *Soc/Geo Groupings*. The remaining 17 results, from four papers, deal with other types of natural identity, which cannot appropriately be fitted into the above categories. These observations relate to political identity (Abbink and Harris, 2012), disability (Gneezy et al., 2012), caste (Hoff et al., 2011) and whether farmers are private or members of cooperatives (Hopfensitz and Miquel-Florensa, 2013). We pool them under the composite variable *Natural Other*. [11,12]

2.2.3. Other variables

In our regressions we include as a dummy variable (*Students*) whether each observation derives from a sample consisting predominantly of students or non-students. Even if not explicitly stated, we assume experiments run at universities have at most a very small number of non-student participants. Likewise, while we accept experiments in the field may include a few student subjects, their proportion is likely to be low (unless otherwise stated). As another control, we include the size of the active decision-making sample from which a given result is derived (*Sample Size*).

[9] We ran further regressions in which we categorised the role types differently. In these models, dummy variables were assigned to specific game settings, such as the trust game sender and the trust game returner. The results are reported in Appendix B.

[10] There is some inconsistency in the literature on the definition of 'minimal groups'; some authors (e.g. Chen and Chen, 2011) categorise certain artificial groups as 'near minimal'. For our purposes, we use 'minimal groups' synonymously with 'artificially created groups.' In Appendix B, we explore the effects of inducing artificial identity using different methods, and show that it seems not to matter precisely how 'minimal' the groups are.

[11] The distinction between *Soc/Geo Groupings* and *Natural Other* is not arbitrary: in- and out-groups in the *Natural Other* category are not necessarily socially or geographically distinct. However, if the *Natural Other* observations are incorporated into the *Soc/Geo Groupings* category, the *Soc/Geo Groupings* coefficients do not change substantially and all other results discussed in the paper remain unaffected.

[12] Two papers provide separate results on more than one natural identity category.

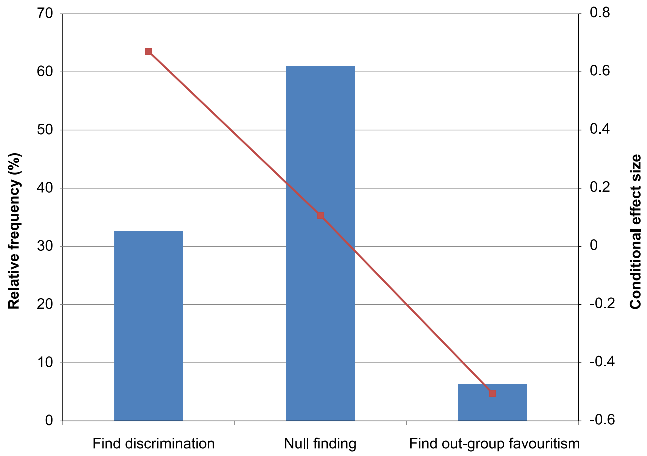

Fig. 1. reakdown of data-points by result type. Note: Blue bars show the percentage of observations in our dataset which find significant discrimination (at the 5% level), a null result, and significant out-group favouritism (at the 5% level). Red points show the average effect sizes for observations belonging to each category. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Breakdown of data-points by result type and identity type.

Category	Obs.	Find discrimination (%)	Find null (%)	Find out-group favouritism (%)	Obs. with available effect sizes	Average Effect size (d) (with 95% C.I. below)
Artificial	169	42.0	55.6	2.4	150	0.365(0.279–0.450)
Natural						
National	82	18.3	68.3	13.4	52	0.164 (0.042–0.286)
Ethnic	63	11.1	82.6	6.3	59	0.134 (0.013–0.255)
Gender	32	9.4	65.6	25.0	28	−0.177 (−0.301 to −0.053)
Religious	21	14.3	80.9	4.8	21	0.034(−0.062 to 0.131)
Soc/Geo	57	64.9	35.1	0.0	51	0.551 (0.432–0.669)
Natural Other Groupings	17	47.1	52.9	0.0	7	−0.036(−0.158 to 0.086)

Notes: For each identity type: the number of observations in our dataset; the percentage of these observations that find significant discrimination (at the 5% level), null results, and significant out-group favouritism (at the 5% level); the number of observations for which effect sizes are calculable; and the weighted average effect size across such observations, with associated 95% confidence intervals.

3. Results

3.1. What is the general pattern of discrimination across the literature?

In total, as shown in Fig. 1, there are 144 results indicating significant discrimination (32.65%), 28 indicating significant out-group favouritism (6.35%), and 269 indicating no significant discrimination or out-group favouritism (61.00%). 57 of our 77 studies record at least one result of discrimination, while only 15 record any results of out-group favouritism. 10 studies separately record results of discrimination and out-group favouritism. The general tendency, then, leans towards insignificant results, although only 15 studies consist entirely of nulls.

For the sub-sample where we are able to generate effect sizes (364 of 441 observations), the random effects meta-analysis finds an overall effect size of 0.256 (95% confidence range: 0.209–0.304). This can be interpreted as, on average, subjects' discriminating against out-groups by about a quarter of a standard deviation. This is not significantly different from Balliet et al. (2014), who find an overall effect size of 0.32 (95% confidence range: 0.27–0.38). Fig. 1 also displays point estimates for aggregate effect sizes, conditional on the type of result found for each observation. Observations finding significant discrimination have an average effect size of 0.67, those yielding null results have an average effect size of 0.11, and those finding significant out-group favouritism have an average effect size of −0.51; this confirms that the strength of the effect size tends to be closely related to the type of result found for a given observation.

Table 2a
Linear probability regressions on discrimination and meta-regressions on effect size.

Dependent variable	Discrimination		d
	LPMa	LPMb	Metareg
Identity			
Ethnicity	−0.293***	−0.294***	−0.140*
	(0.067)	(0.070)	(0.079)
Religion	−0.235*	−0.256*	−0.164
	(0.131)	(0.144)	(0.125)
Nationality	−0.240***	−0.163	−0.145*
	(0.079)	(0.099)	(0.075)
Gender	−0.312***	−0.335***	−0.456***
	(0.068)	(0.072)	(0.099)
Soc/Geo Groupings	0.252**	0.229*	0.354***
	(0.099)	(0.122)	(0.089)
Natural Other	−0.056	−0.243	−0.236
	(0.165)	(0.197)	(0.192)
Role Types			
First Mover	−0.033	−0.025	0.136*
	(0.074)	(0.081)	(0.079)
Second Mover	−0.079	−0.066	0.050
	(0.065)	(0.075)	(0.085)
Simultaneous Mover	0.015	0.023	0.095
	(0.102)	(0.117)	(0.095)
Allocator	0.371***	0.408***	1.077***
	(0.094)	(0.140)	(0.155)
Partner Chooser	0.070	0.118	0.110
	(0.108)	(0.113)	(0.127)
Controls			
Students	0.005	−0.025	0.086
	(0.063)	(0.076)	(0.077)
Sample Size	$6.6e^{-4}$	$4.9e^{-4}$	$6.6e^{-4}$
	$(4.8e^{-4})$	$(7.2e^{-4})$	$(4.8e^{-4})$
Constant	0.406***	0.422***	0.144*
	(0.089)	(0.104)	(0.105)
R^2 (adjusted in Metareg)	0.201	0.196	0.240
N	441	364	364

Notes: LPMa is linear probability model run on full sample, Metareg is meta-regression run on sample for which effect sizes are available, LPMb is linear probability model run on same sample as Metareg; omitted categories are Dictator (role type) and Artificial (identity); errors in LPM models are corrected for heteroskedasticity, with 77 clusters in LPMa and 67 in LPMb; standard errors in italics.

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Result 1. : In general, there is limited discrimination against the out-group.

3.2. How does the level of discrimination vary according to the type of identity groups are based upon?

Table 1 displays a breakdown of our sample's observations by identity category, and the results of random effects meta-analyses run on these sub-samples. For most categories the tendency is towards null results. Only for *Soc/Geo Groupings* – which yields no results of out-group favouritism – are observations of discrimination more likely than insignificant results, and this is also the identity type with the highest average effect size. The category for which there is least discrimination and most out-group favouritism is gender; the average effect size for this sub-sample is negative.

Table 2a extends the analysis of Table 1 through the use of regressions. LPMa is a linear probability model with the dependent variable discrimination against the out-group (equal to 1 if discrimination is found, 0 otherwise). Metareg is a meta-regression with the dependent variable the discrimination effect size. In both models artificial identity and the

Table 2b
Linear Restriction Tests on models presented in Table 2a.

	Null Hypothesis	*P* value on two-tailed test		
		LPMa	LPMb	Metareg
Identity				
	Ethnicity = Religion	0.662	0.776	0.848
	Ethnicity = Nationality	0.533	0.201	0.952
	Ethnicity = Gender	0.785	0.556	0.007***
	Ethnicity = Soc/Geo Groupings	< 0.001***	< 0.001***	< 0.001***
	Ethnicity = Natural Other	0.144	0.786	0.612
	Religion = Nationality	0.973	0.545	0.889
	Religion = Gender	0.573	0.582	0.059*
	Religion = Soc/Geo Groupings	< 0.001***	0.001***	< 0.001***
	Religion = Natural Other	0.365	0.951	0.719
	Nationality = Gender	0.341	0.033**	0.006***
	Nationality = Soc/Geo Groupings	< 0.001***	0.005***	< 0.001***
	Nationality = Natural Other	0.294	0.699	0.648
	Gender = Soc/Geo Groupings	< 0.001***	< 0.001***	< 0.001***
	Gender = Natural Other	0.134	0.64	0.297
	Soc/Geo Groupings = Natural Other	0.080*	0.018**	0.002***
Role Types				
	First Mover = Second Mover	0.444	0.552	0.175
	First Mover = Simultaneous Mover	0.617	0.639	0.589
	First Mover = Allocator	< 0.001***	0.004***	< 0.001***
	First Mover = Partner Chooser	0.309	0.17	0.831
	Second Mover = Simultaneous Mover	0.329	0.399	0.579
	Second Mover = Allocator	< 0.001***	0.001***	< 0.001***
	Second Mover = Partner Chooser	0.129	0.082*	0.621
	Simultaneous Mover = Allocator	0.004***	0.011**	< 0.001***
	Simultaneous Mover = Partner Chooser	0.636	0.434	0.902
	Allocator = Partner Chooser	0.021**	0.064*	< 0.001***

Note: LPMa is linear probability model run on full sample, Metareg is meta-regression run on sample for which effect sizes are available, LPMb is linear probability model run on same sample as Metareg.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

dictator game are the benchmark categories.[13] In these regressions we test whether our identity-type variables still yield significantly different levels of discrimination after controlling for other factors. Table 2b presents the results of linear restriction tests run on the sets of dummy variables featuring in the models.

In both the linear probability models and meta-regression, the identity category linked to the strongest discrimination is social and geographical groupings. In Metareg it yields significantly higher discrimination, at the 1% level, than any of the other identity categories. In LPMa it does the same, except that the differences with *Artifical* and *Natural* Other are only significant at the 5% and 10% level respectively.

The identity category linked to the weakest discrimination is gender. Both the linear probability model and the meta-regression indicate weaker discrimination between genders than between artificial groups, significant at the 1% level. The meta-regression also finds gender discrimination to be weaker than ethnic and national discrimination (at the 1% level), and religious discrimination (at the 10% level). However, LPMa does not find these differences to be significant.[14]

In LPMa the coefficients on the ethnic and national identity types are significantly negative at the 1% level, strongly indicating that discrimination is less likely to be observed when subject pools are split along these lines than on the basis of

[13] By necessity, the choices of omitted categories are somewhat arbitrary – there are no variables to serve as obvious baselines for comparison. We selected *Artificial* because we regard comparisons between discrimination with artificial and natural forms of identity to be particularly interesting (as discussed in Section 1). For role type; we selected *Dictator* because it is a commonly used game in experimental economics and arguably the simplest, making it a useful object for comparison.

[14] *Gender* is also found to be significant in a linear probability regression with out-group favouritism as the dependent variable, presented as LPMa1 in Appendix C, Table C.1. Our results show that gender experiments are more likely to yield observations of out-group favouritism than all other identity types except *Nationality*, with all differences significant at the 1% level. This model additionally finds experiments with socially or geographically distinct groups are less likely to provide results of out-group favouritism than those with artificial or national groups. Other identity types are not associated with significantly strong or weak out-group favouritism – however, we have few results of out-group favouritism across our sample. Where we do find significant identity type effects on out-group favouritism, they are in directions consistent with the results on discrimination – when an identity type is positively (negatively) associated with out-group favouritism, it will be negatively (positively) associated with discrimination.

artificial identities. According to Metareg, however, ethnic and national identity experiments are only linked to significantly lower discrimination (i.e. less positive effect sizes) than artificial group experiments at the 10% level.

Given the inconsistency, Table 2a also reports LPMb, a linear probability model run on the reduced sample for which effect size calculation is possible. This helps to distinguish whether the losses of significance when moving from LPMa to Metareg are due to the reduction in sample or the change in measurement technique. For the comparison of national and artificial identity, the loss of significance appears to be due to the change in sample, as in LPMb the coefficient is also insignificant. The same cannot be said for *Ethnicity*, however, as the linear probability model on the reduced sample continues to report significantly less discrimination between ethnic than artificial groups at the 1% level. Doubt, therefore, is cast over the robustness of our finding on ethnicity – although the coefficient's sign is at least weakly significant.[15,16]

Result 2. The strength of discrimination depends upon the type of group identity under investigation. It is stronger when identity is artificially induced in the laboratory than when the subject pool is divided by ethnicity or nationality, and higher still when participants are split into socially or geographically distinct groups.

3.3. How does the level of discrimination depend upon the decision-making context?

Inspection of the coefficients on role type dummies in LPMa and Metareg (Table 2a) reveals discrimination is significantly stronger when the decision-maker is a third-party allocator than when he or she is a dictator (the omitted category). Linear restriction tests (Table 2b) also show the third-party allocator role is more likely to be associated with discrimination than all the other role types, with the difference always significant at the 1% level under both models. The size of the *Allocator* coefficients in the meta-regression (1.077) is worth noting – it indicates that discrimination in games of this type tends to be very large indeed, with on average more than one standard deviation between subjects' treatment of in- and out-groups.

The other role types do not carry significantly different effects from one another. This is at odds with Kiyonari and Yamagishi (2004) and Balliet et al. (2014), who find discrimination to be stronger by simultaneous movers than first movers (Balliet et al. do not investigate second movers). In an attempt to discern why our result differs from that of Balliet et al., we re-ran our analysis keeping only the observations included in their study. We found there was still no significant difference between *First Mover* and *Simultaneous Mover* (the remaining sample on which to run this regression was small; however, we also compared the aggregate effect sizes for each category and found they are very similar). This suggests the significance of the finding in Balliet et al. is driven by studies outside our dataset, i.e. outside the economics literature.[17]

Result 3. Third-party allocators discriminate more than decision-makers in all other roles.

3.4. Do students discriminate any more or less than non-students?

Most decision-makers in our analysis were students. Only 101 observations, from 22 studies, are produced by in-groups not comprised (at least in their near-entirely) of university students. 31.8% of the observations for students return discrimination, while 6.8% find out-group favouritism and 61.5% are null; for non-students 35.6% find discrimination, 5.0% yield out-group favouritism and 59.4% are null. The coefficient on *Students* is not significant in any of our regressions. That experiments with students do not generate significantly different levels of discrimination than those with non-students is an interesting non-result which suggests that, in this literature, working with student samples will not generate a biased perception of the extent and magnitude of discrimination by the wider population.[18]

Result 4. : Discrimination does not significantly differ between students and non-students.

3.5. Does the experimental literature provide more support for taste-based or statistical theories of discrimination?

For 262 (59.4%) of our observations, as a result of the experimental design any discrimination must be taste-based, as it cannot be statistical. Statistical discrimination cannot occur when a player is making the only or last move in a game, unless

[15] In Table 3, we will later present a meta-regression with the number of role type dummies reduced from five to one. The purpose of this model is to investigate taste-based and statistical discrimination. However, it is worth noting that in this model with fewer independent variables, the coefficient on *Ethnicity* is found to be significantly negative at the 5% level. This improves our confidence that there is an effect. The coefficient on *Nationality* is also significant (at the 1% level) in that model.

[16] We are particularly interested in the finding that discrimination is stronger in artificial group experiments than those employing certain types of natural identity. In an attempt to gain a greater understanding of what drives discrimination between artificial groups, we ran regressions focusing on just the artificial identity sample, coding for the method experimenters used to create artificial groups. We find it makes no difference whether groups are based on preferences (such as for a particular painting) or sheer randomisation. Furthermore, we do not find that team-building exercises designed to strengthen artificial group identity significantly increase the level of discrimination. These results are all presented in greater detail in Appendix B.

[17] With out-group favouritism as the dependent variable (LPMa1 in Appendix C, Table C.1), we find no significant differences at all between any role type pair.

[18] In Appendix B, we also show that the country where an experiment is run is not a significant predictor of the extent of the discrimination found.

Table 3
Linear probability regressions on discrimination and out-group favouritism, and meta-regression on effect size, with or without scope for statistical discrimination.[23]

Dependent variable	Discrimination LPMa	D Metareg
Type of discrimination possible		
Taste+Statistical	0.009	0.071
	(0.056)	(0.053)
Identity		
Ethnicity	−0.285***	−0.189**
	(0.063)	(0.080)
Religion	−0.279**	−0.232*
	(0.134)	(0.131)
Nationality	−0.237***	−0.225***
	(0.072)	(0.079)
Gender	−0.315***	−0.545***
	(0.064)	(0.100)
Soc/GeoGroupings	0.238**	0.265***
	(0.097)	(0.093)
Natural Other	0.074	−0.306
	(0.266)	(0.202)
Controls		
Students	0.043	0.116
	(0.072)	(0.080)
Sample Size	$4.5e^{-4}$	$-2.7e^{-4}$
	$(4.7e^{-3})$	$(4.4e^{-3})$
Constant	0.350***	0.238**
	(0.089)	(0.093)
R^2 **(adjusted in Metareg)**	0.157	0.138 (adjusted)
N	441	364

Notes: LPMa is linear probability model run on full sample, Metareg is meta-regression run on sample for which effect sizes are available; omitted categories are taste-based only (type of discrimination possible) and Artificial (identity); errors in LPMa are corrected for heteroskedasticity, with 77 clusters; standard errors in parentheses.
* $p<0.1$.
** $p<0.05$.
*** $p<0.01$.

the game is to be repeated, or possibly if the move is made simultaneously with others. Discrimination by trust game returners, for example, can only be taste-based, because opponents then have no control over the final outcome and beliefs about their type are therefore irrelevant.[19] All observations under the *Dictator* and *Allocator* categories preclude the possibility of statistical discrimination, as do all except seven (due to the game being repeated) in the *Second Mover* category. All observations under the *First Mover* category permit the possibility of statistical discrimination, as do most in the *Partner Chooser* category and around a third in the *Simultaneous Mover* category.[20]

In Table 3, we run a linear probability regression on discrimination and a meta-regression on the discrimination effect size, with role types re-coded into two types: one, *Taste+Statistical*, where there is scope for both taste-based and statistical discrimination, the other (the omitted category) where there is scope only for taste-based discrimination. Note that in this literature any game-role contains scope for taste-based discrimination. The coefficient on *Taste+Statistical* is positive but insignificant in both the linear probability regression ($p=0.87$) and the meta-regression ($p=0.18$). This indicates there is no significant difference in the likelihood of observing discrimination, or in the predicted effect size, when scope for statistical discrimination is added.[21]

[19] There is a grey area to be acknowledged here. One could have a model of statistical taste-based discrimination, in which people have a taste for discrimination against a group because of beliefs they hold about its members (for instance, about how rich they are). In this paper, we do not distinguish between this and any other type of taste-based discrimination (i.e. we do not consider root motivations for taste-based discrimination).

[20] In Appendix A, Table A.2, we list which specific games permit which forms of discrimination.

[21] We also ran a linear probability model on out-group favouritism, with the equivalent specification to LPMa1 in Table 3. This is reported as LPMa2 in Appendix C, Table C.1. As with discrimination, there is no significant difference in the likelihood of observing out-group favouritism when scope for statistical discrimination is added.

This would suggest taste-based discrimination is an important driver of behaviour in these experiments and statistical discrimination is not, but we probe further by analysing the results of individual experiments. Where there is scope for statistically-motivated discrimination, by design for 66.5% of these observations it is not possible to disentangle its effects from taste-based motivations. To be able to do so, an experiment must either use belief elicitation or include a control game in which behaviour can only be taste-based – the most common case of this is adding a dictator game to extricate taste-based from statistical discrimination by trust game senders.[22] In the 60 cases that it is possible to distinguish between discriminatory motives, the authors find significant statistical discrimination to occur in 13 cases (10 times against the out-group and three times in favour of it). Within the same sample, for given beliefs or behaviour in a game with a belief-based component, they find significant taste-based deviations from own-payoff-maximisation in 26 cases (16 times against the out-group and 10 times in favour of it). In 26 cases neither statistical nor taste-based discrimination is found at the 5% level. We list all significant findings of taste-based and statistical discrimination from experiments designed to distinguish between the two in Appendix A, Table A.3.

Although the sample is small, tastes are found to affect behaviour more often than statistical beliefs. It seems, however, that beliefs do play some role in determining discriminatory behaviour in economics experiments. We conjecture that the insignificant regression results in Table 3 may be due to the fact that beliefs can either increase or reduce discrimination. This would be because individuals have favourable beliefs about the cooperativeness of out-groups, or because unfavourable beliefs about the out-group's cooperativeness can in some cases actually lead to statistical out-group favouritism. That is, depending on the game setting, self-serving optimal behaviour can either become more or less generous in response to the perception that one's partner is relatively uncooperative. In ultimatum games, for instance, if proposers expect out-group responders to treat them less favourably than in-group responders do, the self-serving optimum is to send them relatively kind offers. This is in contrast to how first mover behaviour would work in trust games, say, where a self-serving sender will send relatively low investments to an out-group responder if it expects to be treated unfavourably by them.[24,25]

Result 5. : There is evidence for both taste-based and statistical discrimination. Tastes appear to drive the general tendency for discrimination against the out-group, but individual studies have found beliefs to affect discrimination.

3.6. *In gender experiments, how does male-to-female discrimination compare with female-to-male discrimination?*

An immediately obvious finding is that gender acts very differently from other identity types. It is the only identity category which is more likely to be associated with a bias against the in-group than against the out-group, with eight results of the former and three of the latter out of a total 32 observations. On the reduced sample, the random effects meta-analysis finds an overall discrimination effect size of -0.177 (95% confidence range: -0.301 to -0.053) for gender experiments, representing significant out-group favouritism. There is obvious intuition why gender is different from the other identity categories: it is the only case in which the effects of sexual attraction – towards the out-group more than the in-group, for most subjects – and 'chivalry' (Eckel and Grossman, 2001) can be expected.

Every experiment on gender in the meta-analysis has a symmetrical male-female design, meaning that for every estimate of discrimination by men against women there is an identical treatment measuring discrimination by women against men. This allows a very clean comparison of these two behaviours across the sample. The only three significant results in our dataset of one gender discriminating against the other are female decision-makers discriminating against males, while six of the eight significant results of one gender favouring the other are male decision-makers favouring females. However, the calculated overall effect size for female decision-makers is actually slightly more negative than for males: -0.181 (95% confidence range: -0.35 to -0.013) for females and -0.173 (95% confidence range: -0.369 to 0.024) for males, although the difference is far from significant. Note that while the effect size indicates females significantly favour males at the 5% level, the equivalent effect for male decision-makers is only significant at the 10% level.

Result 6. : There is significant out-group favouritism in gender experiments. Females significantly favour males; males favour females but the effect is only weakly significant.

[22] There are no precisely standard methods for disentangling taste-based and statistical discrimination. When using a control game in which only taste-based discrimination is possible, statistical discrimination is identified if this game finds significantly weaker discrimination than the setting with scope for both types of discrimination. When using belief elicitation, statistical discrimination is confirmed if beliefs about the in-group and out-group significantly differ, and there is significant discrimination in the direction that would maximise the decision-makers' payoffs based on these beliefs; taste-based discrimination is confirmed if there is still significant discrimination after controlling for the beliefs. Some studies use regression analysis, others non-parametric tests.

[23] Table 3 does not present an LPMb model because in this case we are not interested in investigating any disparities between LPMa and Metareg – the Taste+Statistical coefficient is insignificant in both models.

[24] We are unable to explore this empirically. We can separate games into those where favourable beliefs about a partner's cooperativeness should either increase or decrease the selfish decision-maker's cooperation towards them, but we would need data on beliefs about in-groups and out-groups to predict the direction of discrimination this should result in.

[25] In Appendix B, section B.4, we analyse how the relative strength of discrimination in experiments featuring different identity types interacts with the type of discrimination possible. We show discrimination is only significantly stronger across artificial groups than across ethnic, religious or national groups when there is no scope for statistical discrimination, while discrimination is only significantly stronger across social/geographical groups than across artificial groups when there is scope for statistical discrimination.

4. Discussion and conclusions

A leading result of this paper is that discrimination in economics experiments varies by the type of identity groups are based upon. It is very strong when groups are socially or geographically distinct, and is relatively weak when they are based on ethnicity or nationality. Notably, it tends to be relatively strong in experiments using artificially-induced group identities – so it can confidently be stated that minimal groups do not produce the minimal level of discrimination. At first glance, this seems surprising.

It might be that artificial group manipulations are stronger priming instruments than natural identity experiments tend to use – after all, these dedicate an entire preliminary phase of the experiment to inducing feelings of identity, which will remain at the front of subjects' minds when they are then offered the chance to discriminate. This explanation is arguably supported by the evidence of Robbins and Krueger (2005), whose meta-analysis of psychology experiments shows subjects exhibit stronger in-group projection – that is, they perceive in-group members to be particularly similar to them, relative to out-group members – when identities are artificial than when they are natural. On the other hand, we do not find that team-building exercises, which are designed specifically to strengthen artificially-induced identity and would seem to amplify priming, have a significant effect on the level of discrimination (this is consistent with the findings of Chen and Li (2009)).

Conversely, it could be argued that, for the populations studied in the literature, membership of particular ethnic and national groups does not actually instil strong identity, so that even such trivial identities as can be artificially induced have a greater effect. There is evidence that the process of globalisation has weakened national and ethnic parochialism (Buchan et al., 2009), and in recent decades youth identity in the West and increasingly elsewhere has come to define itself to a large extent upon individuals' belonging to subcultures based on fashion and music tastes – preferences drawn from choice sets which are not, indeed, so different from the apparently arbitrary minimal group painting dichotomy. However, it would seem highly complacent to draw the conclusion from our results that racism and xenophobia are not big problems in many societies.

Another explanation may be that subjects in ethnic and national identity experiments are shying away from displaying 'politically incorrect'[26] behaviour, given that racism and xenophobia are taboo in most societies today. While the link between social acceptability and discrimination has not been well explored, the prejudice literature has yielded relevant findings: that expressions of prejudice correlate with perceptions towards the social acceptability of such prejudice (e.g. Crandall et al., 2002), and furthermore that this correlation is at least partly the result of norm-compliance (e.g. Blanchard et al., 1994).

It seems unlikely that discriminating on the basis of a stated preference for Klee's paintings over Kandinsky's carries any taboo similar to ethnic or national discrimination. Indeed, some subjects may regard an artificial group situation as a game in which they belong to one of the teams, wherein the social norm actively encourages favouritism of one's own group – the sheer strangeness of the setting may even lead subjects to perceive a demand for discrimination on the part of the experimenter (see e.g. Zizzo (2010)). Concerns about social acceptability could explain also why the *Soc/Geo Groupings* category produces significantly higher discrimination than other types of natural identity. Of course, it would not be surprising if relational and geographic proximity led to a stronger sense of belonging than shared ethnicity, religion or nationality, but bear in mind too that there is arguably no taboo against favouring friends over strangers.[27]

If it were shown that discrimination in economics experiments is indeed limited by concerns about social acceptability, it might cast doubt over the external applicability of such studies' findings. It is possible that if participants guess an experiment is about a type of discrimination which is taboo, it will systematically generate a lower effect than if the subjects were unaware of its purpose. On the other hand, the very same concerns about social acceptability might also limit certain types of discrimination outside the lab.

It is noteworthy that gender is the identity category producing the weakest discrimination: in fact, here the meta-analysis finds a significant amount of out-group favouritism. However, gender discrimination clearly persists in the outside world. It may be that economics experiments do not find it because they poorly reflect the conditions under which it survives beyond the lab – in particular, in the job market.

It would be interesting to see more experiments designed to directly compare the effects of different types of group identity. This meta-analysis includes just four. Dugar and Shahriar (2009), Li et al. (2011) and Goette et al. (2012a) all compare discrimination between social/geographical groups and artificial groups, while Abbink and Harris (2012) use artificial groups and political groups (which fall under the *Natural Other* category). The results of all four studies are consistent with ours – discrimination is always lower with artificial identity. However, direct comparisons between artificial group and ethnic or national discrimination are lacking, and it would be very illuminating to see whether such studies support – and if so, whether they can explain – the findings of this meta-analysis.

What implications does our research have for future experiments on discrimination? First, using artificially induced identities as a control against which to pit the results of natural identity treatments may not be recommendable, as the artificial group manipulation appears not so much to capture the minimal level of discrimination that must result from priming any type of identity in a laboratory as to in fact often go beyond that.

Regarding role type, we find discrimination by third-party allocators is much stronger than by participants in any other game setting. If social acceptability does indeed limit discrimination, this is a counterintuitive result, as the allocator role essentially invites subjects to overtly and consciously favour one group over another and therefore seems to be the one that

[26] Political correctness is defined as 'The avoidance of forms of expression or action that are perceived to exclude, marginalize, or insult groups of people who are socially disadvantaged or discriminated against' (Oxford Dictionaries).

[27] This does depend upon the context, however. There are strong taboos against nepotism in certain labour-market transactions. Possibly, the experiments in this literature do not recreate such circumstances.

most obviously telegraphs the purpose of this type of experiment. One possibility is that the role carries an experimenter demand effect – whereby subjects feel they are encouraged to discriminate – or even an action bias effect, if the equal split feels like a default non-move. Another relevant factor may be that the third-party allocator is unique amongst our role types in the decision-maker's payoff being entirely disconnected from the extent to which they discriminate. In any case, experimenters should bear in mind that because they are more likely to identify significant discrimination when they employ the allocator role, they should be less confident that the same groups will discriminate against each other in different contexts.

We find the strength of discrimination does not significantly differ between student and non-student subject pools. This suggests – unlike in the context of social preferences (e.g. Bellemare and Kroger, 2007; Anderson et al., 2013) – student subjects are not a generally unrepresentative sample for questions relating to discrimination. However, we do not exclude the possibility that they are unrepresentative in specific instances, or within particular societies.

There is scope for more experimental research investigating taste-based and statistical discrimination. We show both are relevant, and the two types manifest themselves to different extents in different contexts. However, relatively few experiments have been designed to distinguish between taste-based and statistical discrimination, and more could be known about the mechanisms underlying them.

As a final observation, there is a great deal of variation in the findings of the experimental economics discrimination literature. Our analysis can explain some of it, but our LPM regressions typically have R^2 statistics below 0.2, and the meta-regressions' Adjusted R^2s are rarely above 0.35. As might be expected, discrimination does seem to vary idiosyncratically and is not easy to predict. The results of natural identity experiments do not seem very generalizable – they probably reflect more the characteristics of the specific groups under investigation, and the relationships between them, than aspects of the experimental design. Whilst a drawback for some research questions, this also means there is a great deal of scope for future experimental studies aimed at measuring the levels of discrimination within subject pools of specific interest. Furthermore, given the potential concerns we raise about experimenter demand effects and the external validity of lab experiments on discrimination, the important role of field experiments should be emphasised. Subjects in such studies are unaware they are being observed by experimenters and their behaviour can therefore not be influenced by the fact. Field experiments can test the generalisability of lab findings on discrimination.

Acknowledgements

Many thanks to Abigail Barr and Daniele Nosenzo for their support and input. I am very grateful to all authors who sent data for this project. I received helpful comments at presentations in Nottingham, Duisburg and Warwick from, amongst others, Swee-Hoon Chuah, Simon Gächter, Gianluca Grimalda, Heike Hennig-Schmidt and Gari Walkowitz. Thanks to the Economic and Social Research Council Award No. ES/K002201/1 for financial support.

Appendix A

See Table A1–A3.

Table A.1
Subjective decisions on appropriately defined groups.

Study	Notes
Burns (2004)	We consider 'coloured' to be an appropriate ethnic group in South Africa (as defined in comparison to 'white' and 'black').
Chen et al. (2014)	We do not consider 'Asian' to be an appropriate ethnic group in the USA.
Ferraro and Cummings (2007)	We consider 'Hispanic' to be an appropriate ethnic group in the USA. (Justification, relative to 'Asian': Hispanic people in the USA share a more unified culture than those of Asian descent; they are descended from more linguistically homogeneous peoples than Asians)
Friesen et al. (2012)	We do not consider 'East Asian' and 'South Asian' to be appropriate ethnic groups in Canada.

Table A.2
Game types in meta-analysis and how they measure discrimination.

Role Type	How discrimination is measured	Type of discrimination possible
Trust Game Returner	Difference in proportion of amount received from the sender that is returned*, between in-group and out-group matching. *NB: Ploner and Soraperra (2004) use Indirect Trust Game, where the amount returned is not given to the sender but a group member of theirs	Taste-based only
Agent in Principal-Agent Game (Masella et al., 2012)	Difference in amount sent to principal, between in-group and out-group matching.	Taste-based only
Dictator; Proposer in Unilateral Power Game (Zizzo, 2011)	Difference in amount sent to recipient, between in-group and out-group matching. (NB: Büchner and Dittrich (2002) use a saving game where one player leaves the game early and decides how much to leave their partner. This decision is the equivalent of that faced by a dictator)	Taste-based only
Allocator	Difference in amount allocated to in-group and out-group member.	Taste-based only
Responder in Ultimatum Game, Hold-up Game (Morita and Servátka, 2013)	Difference in likelihood of rejecting an offer, controlling for its size, between in-group and out-group matching.	Taste-based only
Responder in Proposer-Responder Game (McLeish and Oxoby, 2007)	Difference in amount by which proposer's payoff is reduced, controlling for amount offered by proposer, between in-group and out-group matching.	Taste-based only
Responder in Proposer-Responder Game (Chen and Li, 2009; Currarini and Mengel, 2012)	Difference in rate of choosing more other-regarding response, between in-group and out-group matching.	Taste-based only
Third-party punisher	Difference in punishment level, controlling for behaviour of punishee, between in-group and out-group matching.	Taste-based only
One-shot Prisoner's Dilemma	Difference in rate of cooperation, between in-group and out-group matching.	Taste-based only
Trader in market games	For bidders: difference in price offered, between in-group and out-group matching. For sellers: difference in price accepted, between in-group and out-group matching.	Taste-based only for sellers in one-shot interactions; taste-based and statistical for bidders in one-shot interactions, and for all players in repeated games.
Public Goods Game	Difference in contribution level, between in-group and out-group matching.	Taste-based only in one-shot games (Hopfensitz, 2013); taste-based and statistical in repeated games.
Partner-Choosing Role	Difference in rate of choosing in-group partner and out-group partner.	Taste-based only if chosen partner does not become active decision-maker in subsequent games; taste-based and statistical if they do.
Repeated Common Pool Withdrawal Game Carpenter and Cardenas, (2011)	Difference in withdrawal level, between in-group and out-group matching.	Taste-based and statistical.
Minimal Effort Game (Chen and Chen, 2011)	Difference in effort level, between in-group and out-group matching.	Taste-based and statistical.
Trust Game Sender; Principal in Principal-Agent Game (Masella et al., 2012); First Mover in Hold-up Game (Morita and Servatka, 2013)	With continuous action space: difference in amount sent, between in-group and out-group matching. With binary action space: difference in rate of choosing to trust, between in-group and out-group matching.	Taste-based and statistical.
Investor in Investment Game (Wu, 2009)	Difference in amount invested in manager's project, between in-group and out-group matching.	Taste-based and statistical.
Ultimatum Game Proposer; Second Mover in Hold-up Game (Morita and Servatka, 2013); First Mover in Proposer-Responder Game (McLeish and Oxoby, 2007)	Difference in amount offered, between in-group and out-group matching.	Taste-based and statistical.
Proposer in Proposer-Responder Game (Chen and Li, 2009; Currarini and Mengel, 2012)	Difference in rate of choosing more other-regarding first move, between in-group and out-group matching.	Taste-based and statistical.
Nash Demand Game (Ruffle and Sosis, 2006; Zizzo, 2011)	Difference in amount claimed, between in-group and out-group matching	Taste-based and statistical.
Stag Hunt	Difference in rate of choosing hawkish strategy, between in-group and out-group matching.	Taste-based and statistical.

Table A.3
List of significant results from studies designed to distinguish between taste-based and statistical discrimination

Result	Paper	Role	Groups
Taste-based and statistical discrimination	Banuri et al. (2012)	Partner-choosing role	Colleges within university (Free Nepotism treatment)
	Bernhard et al. (2006)	Dictator	Tribes
	Binzel and Fehr (2013)	Trust Game Sender	Social groups
	Currarini and Mengel (2012)	Partner-choosing role	Artificial (comparison of COORD, ENDO and LOWB treatments)
	Etang et al. (2011a)	Trust Game Sender	Villages
Taste-based discrimination only	Burns (2004)	Trust Game Sender	Ethnic (coloured in-group, black out-group)
		Trust Game Sender	Ethnic (coloured in-group, white out-group)
	Chuah et al. (2013)	Trust Game Sender	Religious (Hindu in-group, Muslim out-group)
		Trust Game Sender	Religious (Muslim in-group, Hindu out-group)
	Ferraro and Cummings (2007)	Ultimatum Game Proposer	Ethnic (Hispanic in-group, Navajo out-group)
	Guillen and Ji (2011)	Trust Game Sender	National (Australian in-group, non-Australian out-group)
	Kim et al. (2013)	Trust Game Sender	National (North Korean in-group, South Korean out-group – sample 1)
		Trust Game Sender	National (North Korean in-group, South Korean out-group – sample 2)
	McLeish and Oxoby (2007)	First Mover in Proposer-Responder Game	Artificial (OP treatment)
		First Mover in Proposer-Responder Game	Artificial (NO treatment)
	Ruffle and Sosis (2006)	Nash Demand Game	Social/geographical (Kibbutz in-group, city out-group)
Statistical discrimination only	Banuri et al. (2012)	Partner-choosing role	Colleges within university (Costly Nepotism treatment)
	Boarini et al. (2009)	Ultimatum Game Proposer	National (French in-group, Indian out-group)
	Chen and Chen (2011)	Minimal Effort Game	Artificial (Enhanced treatment)
	Haile et al. (2008)	Trust Game Sender	Ethnic (white in-group, black out-group)
	Masella et al. (2014)	Principal in Principal-Agent Game	Artificial
Taste-based out-group favouritism only	Burns (2004)	Trust Game Sender	Ethnic (white in-group, black out-group)
		Trust Game Sender	Ethnic (black in-group, white out-group)
		Trust Game Sender	Ethnic (black in-group, coloured out-group)
	Hennig-Schmidt et. al. (2007)	Trust Game Sender	National (Israeli in-group, Palestinian out-group)
		Trust Game Sender	National (Palestinian in-group, Israeli out-group)
	Kim et al. (2013)	Trust Game Sender	National (South Korean in-group, North Korean out-group – sample 1)
		Trust Game Sender	National (South Korean in-group, North Korean out-group – sample 2)
	Slonim and Guillen (2010)	Trust Game Sender	Gender (male in-group, Gender/Ability Selection treatment)
		Trust Game Sender	Gender (male in-group, No Selection treatment)
		Partner-choosing role	Gender (male in-group, Trust Game treatment)
Statistical out-group favouritism only	Boarini et al. (2009)	Ultimatum Game Proposer	National (Indian in-group, French out-group)
	Hennig-Schmidt et al. (2007)	Trust Game Sender	National (German in-group, Palestinian out-group)
	Slonim and Guillen (2010)	Partner-choosing role	Gender (female in-group, Trust Game treatment)

Appendix B. : Further results

B.1 Further analysis of role types

In this section we investigate the effects on discrimination of using specific game types. We recode the role type variables, assigning dummies to specific game settings in the following way. Trust games and similar principal-agent games provide two roles: senders (*TG Sender*, 98 observations) and returners (*TG Returner*, 81). The next most common role type is the *Dictator* (68). Prisoner's dilemmas, public goods games, and common pool withdrawal games are all social dilemmas, and are coded under a single category (*Social Dilemma*, 58). Next we have third-party allocators (*Allocator*, 33). Ultimatum games and similar bargaining settings are grouped together and split into two role types: first movers (*Proposer*, 31) and second movers (*Responder*, 27). Treating *Dictator* as the omitted category in our regressions, we form a set of binary independent variables from the other six role types, plus the additional variable *Game Other* (45 observations) into which are placed the remaining game settings that we did not think could be adequately categorised.[28]

Table C.3a in Appendix C displays the output of regressions incorporating these variables. These regressions are the equivalent of those presented in Table 2, the only change being the recoding of the role type variables. As above, LPMa1 and Metareg1 show discrimination to be significantly stronger when the decision-maker is a third-party allocator than when he or she is a dictator (the omitted category). Linear restriction tests (Table C.3b) also show the third-party allocator role is more likely to be associated with discrimination than all the other role types, with the difference always significant at the 1% level under both models. Again, the other role types do not consistently carry significantly different effects from one another. This is at odds with the analysis of Balliet et al. (2014), who find discrimination is stronger by trust game senders than by dictators, and stronger still in social dilemmas. With out-group favouritism as the dependent variable (LPMa2), the only significant differences between role types are that proposers are less likely to engage in out-group favouritism than dictators, trust game senders, trust game returners and subjects in the *Game Other* category.

Result A1. : We do not find strong effects associated with such specific role types as the trust game sender or returner, players in social dilemmas, or bargaining game proposers or responders.

B.2 Does the strength of discrimination in artificial group experiments depend on the method used to induce identity?

The way in which identity is artificially instilled in subjects varies from experiment to experiment. However, we can identity two broad categories of artificial group creation. One follows the original Tajfel et al. (1971) process of allowing subjects to self-select into groups. Typically this involves asking participants to choose a preference between the art of Klee and Kandinsky, although some studies elicit preferences on other choice sets, such as favourite colours. We code these observations under *Preferences*. The other main category gives subjects no control over which group they belong to. In such cases they are simply randomly assigned and labelled as belonging to, for instance, the ‘red’ or ‘blue’ group. We code these manipulations as *Labelling*. Occasionally, a different type of identity inducement is done – for example, groups can be based on subjects’ tendency to overestimate or underestimate the number of dots on a screen (Guala et al., 2013; Ioannou et al., 2013), or by the time at which they undertake a particular task (Ahmed, 2007). These cases we code as *Other Method*.

Another way artificial group manipulations vary is by whether they contain additional stages in which group members interact, between being placed into groups and before the task upon which discrimination is measured. These stages often involve games in which group members must work together to earn monetary rewards, although on some occasions they merely interact non-strategically as a result of being permitted to converse electronically. Such stages are introduced as a mechanism to strengthen artificial group identities. We code their presence in studies under *Team Building*.

In order to test how these different procedures affect the extent of discrimination, we run LPM and meta-regressions on our sub-sample of observations for which identity is artificial. These are presented in Appendix C, Table C.4. We find there is no significant difference between whether groups are self-selected or randomly selected. Also, while the coefficients are in the direction of strengthening discrimination, we find the effect of team-building exercises not to be significant. From these results, we infer that the precise form of identity inducement is not crucial to the outcome of artificial group experiments. This is consistent with the findings of Chen and Li (2009), whose experiment addresses these questions.

Result A2. : The strength of discrimination in artificial group experiments does not depend significantly on the method used to induce identity.

B.3 Can country-level variables explain discrimination?

Our meta-analysis encompasses geographical diversity, with data from 31 countries. Including cases where the out-group was located in a different country, 169 results from 34 studies come from Europe, 116 observations from 22 studies are from North America, 85 results from 17 studies are from Asia, 37 observations from seven studies come from Africa, nine results from three studies come from Latin America, and ten observations from three studies are from Australasia. Ten results from two papers have decision-makers located in more than one country, while one paper does not mention where its experiment took place. The country providing the most observations is the USA, with 106 from 19 studies.

[28] Specifically, the Game Other category consists of players in the following settings: unstructured bargaining games; the battle of the sexes; coordination games; indirect trust games; market-trading games; minimal effort games; Nash Demand games; partner-choosing situations; saving games; stag hunts; and third-party punishment games. Several of these could have been coded under a standalone category – coordination games and variants – but there would only be eight observations in such a category.

This diversity allows us a further set of variables to test for relationships between discrimination and characteristics of the country in which an experiment is run. In Appendix C, Table C.5, therefore, we report regressions including location dummies for the *USA* and *Europe*, and country-level measures of *Individualism* (from the Hofstede Centre), ethno-linguistic-religious *Fractionalisation* (constructed from Alesina et al. (2003), by averaging each country's scores for ethnic, linguistic and religious fractionalization[29]) and prosperity (*Log GDPpc*, the log of per capita national income at purchasing power parity, as estimated by the World Bank). Using these independent variables requires trimming the sample to exclude experiments conducted across countries, as well as those in locations for which data on *Individualism* is not available.

We do not find any country-level variables to be significant, with rare exceptions. In LPMa2, we find the probability of observing out-group favouritism is lower in the USA than in the rest of the world, significant at the 5% level. However, once controlling for country-level individualism, as in LPMa3, the effect disappears. *Individualism* itself only has a weakly significant effect of reducing the likelihood of out-group favouritism, after omitting the USA dummy in LPMa4.

While the insignificance of country-level variables in our analysis appears to show that results on discrimination can be generalised across cultures, we do not argue this is necessarily the case. The locations at which experiments on discrimination have been conducted are not a random global sample; in many cases they are handpicked by researchers who have prior reason to believe they have an interesting discrimination-related question to ask of a particular subject pool.

Result A3. : Country-level variables are not found to significantly explain discrimination.

B.4 How does the experimental context affect the prevalence of each type of discrimination?

To investigate the strength of different types of discrimination in experiments with different types of identity, we run LPM and meta-regressions on the sub-sample of observations for which there is scope only for taste-based discrimination, and the sub-sample for which there is scope for both taste-based and statistical discrimination. The results are presented in Appendix C, Table C.6a; LPMa1 and Metareg1 relate to the taste-based only sub-sample, while LPMa2 and Metareg2 relate to the both-types sub-sample. The table reports whether the coefficients for each identity category significantly differ between models 1 and 2; this is deduced by running pooled models with interaction terms. The results of linear restriction tests are also presented in Appendix C, Table C.6b.

When it can only be driven by taste, according to both the LPM and the meta-regression discrimination is significantly greater across artificial groups than across ethnicities, religions, nationalities or gender. All of these differences are significant at the 1% level, apart from the difference between *Artificial* and *Religion* in Metareg(1), which is significant at the 5% level. However, when discrimination can be driven both by tastes and statistical beliefs, neither model finds it to significantly differ between artificial group experiments and those on nationality, religion or ethnicity.[30]

With only taste-based discrimination possible, discrimination is not significantly different across artificial groups to across socially or geographically distinct groups. However, when there is also scope for statistical discrimination, discrimination is significantly higher (1% level) among socially or geographically distinct groups.

The only identity category whose coefficient significantly differs between the sample where only taste-based discrimination is possible and the sample where both types of discrimination are possible, in models ran on both dependent variables, is *Soc/Geo Groupings*. The coefficients on *Ethnicity*, *Religion* and *Nationality* do not significantly differ between samples. The test on the omitted category, *Artificial*, shows its coefficient also does not significantly differ between samples. We therefore interpret the narrowing of the discrimination gap between *Artificial* and *Ethnicity, Religion* and *Nationality* when scope is added for statistical discrimination as being driven by beliefs either reducing discrimination in artificial identity experiments, or enhancing it in experiments with ethnicity, religion and nationality, or both. We interpret the widening of the discrimination gap between *Artificial* and *Soc/Geo Groupings* when scope is added for statistical discrimination as being driven primarily by beliefs enhancing discrimination between social and geographical groups.

Result A4. : Discrimination is only significantly stronger between artificial groups compared to between ethnic, religious and national groups when there is scope only for taste-based discrimination. Discrimination is only significantly stronger between social/geographical groups compared to between artificial groups when there is scope for both types of discrimination.

Appendix C. : Additional regression output

See Tables C.1,C.2, C.3a,C.3b,C.4,C.5,C.6a,C.6b.

[29] We also ran regressions containing separate variables for ethnic, linguistic and religious fractionalization, none of which were found to have significance.

[30] Except with *Ethnicity* at the 10% level in the LPM.

Table C.1
Linear probability regressions on out-group favouritism.

Dependent variable	Out-group favouritism	
	LPMa1	LPMa2
Type of discrimination possible		
Taste+Statistical		−0.003
		(0.020)
Role Types		
First Mover	−0.031	
	(0.029)	
Second Mover	−0.012	
	(0.027)	
Simultaneous Mover	−0.025	
	(0.038)	
Allocator	−0.049	
	(0.037)	
Partner Chooser	0.031	
	(0.063)	
Identity		
Ethnicity	0.041	0.035
	(0.040)	(0.036)
Religion	−0.004	−0.008
	(0.052)	(0.050)
Nationality	0.118*	0.111
	(0.069)	(0.068)
Gender	0.222***	0.231***
	(0.046)	(0.047)
Soc/Geo Groupings	−0.051**	−0.045*
	(0.024)	(0.023)
Natural Other	−0.025	−0.041
	(0.034)	(0.026)
Controls		
Students	−0.023	−0.026
	(0.041)	(0.038)
Sample Size	$4.9e^{-4}$	$2.0e^{-4}$
	$(7.2e^{-4})$	$(2.9e^{-4})$
Constant	0.051	0.038
	(0.046)	(0.047)
R^2	0.088	0.084
N	441	441

Notes: LPMa1 and LPMa2 are linear probability models run on full sample; omitted categories are Dictator (role type) and Artificial (identity); errors are corrected for heteroskedasticity, with 77 clusters; standard errors in italics.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

Table C.2
Logistic regressions on discrimination and out-group favouritism.

Dependent variable	Discrimination LOGITa1	Out-group favouritism LOGITa2
Identity		
Ethnicity	−0.260***	0.062
	(0.046)	(0.069)
Religion	−0.195**	−0.003
	(0.098)	(0.061)
Nationality	0.216***	0.172*
	(0.065)	(0.095)
Gender	−0.252***	0.326***
	(0.045)	(0.097)
Soc/Geo Groupings	0.244**	(dropped)
	(0.111)	
Natural Other	−0.056	(dropped)
	(0.138)	
Role Types		
First Mover	−0.031	−0.022
	(0.081)	(0.017)
Second Mover	−0.080	−0.007
	(0.066)	(0.017)
Simultaneous Mover	0.025	−0.011
	(0.113)	(0.030)
Allocator	0.413***	−0.035**
	(0.103)	(0.014)
Partner Chooser	0.075	0.035
	(0.125)	(0.045)
Controls		
Students	0.011	−0.037
	(0.072)	(0.053)
Sample Size	$1.7e^{-4}$	$1.9e^{-4}$
	$(4.9e^{-4})$	$(1.8e^{-4})$
Pseudo R^2	0.167	0.132
N	441	367

Notes: LPMa1 and LPMa2 are linear probability models run on full sample; omitted categories are Dictator (role type) and Artificial (identity); errors are corrected for heteroskedasticity, with 77 clusters in LOGITa1 and 66 in LOGITa2; standard errors in parentheses; for dummy variables, dy/dx is for discrete change from 0 to 1.

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Table C.3a
Linear probability regression on discrimination and meta-regression on effect size (further analysis of role type).

Dependent variable	Discrimination		d	Out-group favouritism
	LPMa1	LPMb1	Metareg1	LPMa2
Identity				
Ethnicity	−0.287***	−0.282***	−0.113	0.032
	(0.078)	(0.084)	(0.081)	(0.041)
Religion	−0.219	−0.220	−0.092	−0.027
	(0.141)	(0.156)	(0.128)	(0.057)
Nationality	−0.234***	−0.133	−0.112	0.107
	(0.085)	(0.106)	(0.078)	(0.067)
Gender	−0.314***	−0.339***	−0.500***	0.233***
	(0.065)	(0.069)	(0.098)	(0.042)
Soc/Geo Groupings	0.242**	0.242*	0.365***	−0.057**
	(0.102)	(0.124)	(0.091)	(0.023)
Natural Other	−0.069	−0.286	−0.203	−0.034
	(0.176)	(0.179)	(0.193)	(0.035)
Role Types				
TG Sender	−0.045	−0.045	0.002	$-2.6e^{-4}$
	(0.126)	(0.086)	(0.081)	(0.027)
TG Returner	−0.126*	−0.147*	−0.112	0.016
	(0.074)	(0.083)	(0.087)	(0.024)
Social Dilemma	0.014	0.010	−0.022	−0.011
	(0.106)	(0.115)	(0.097)	(0.039)
Allocator	0.348***	0.400***	0.991***	−0.042
	(0.104)	(0.141)	(0.154)	(0.034)
Proposer	−0.081	−0.131	−0.012	−0.088**
	(0.101)	(0.090)	(0.106)	(0.041)
Responder	−0.029	0.129	0.120	−0.017
	(0.099)	(0.109)	(0.135)	(0.055)
Game Other	0.045	0.087	0.034	0.004
	(0.098)	(0.119)	(0.104)	(0.037)
Controls				
Students	0.004	−0.028	0.103	−0.018
	(0.068)	(0.077)	(0.077)	(0.040)
Sample Size	$1.2e^{-4}$	$5.7e^{-5}$	$-7.1e^{-4}$*	$2.7e^{-4}$
	$(4.3e^{-4})$	$(4.7e^{-3})$	$(4.3e^{-3})$	$(3.2e^{-3})$
Constant	0.416***	0.442***	0.236**	0.037
	(0.087)	(0.100)	(0.106)	(0.048)
R^2 (adjusted in Metareg1)	0.206	0.214	0.237	0.094
N	441	364	364	441

Notes: LPMa1 and LPMa2 are linear probability models run on full sample, Metareg1 is meta-regression run on sample for which effect sizes are available, LPMb1 is linear probability model run on same sample as Metareg1; omitted categories are Dictator (role type) and Artificial (identity); errors in LPM models are corrected for heteroskedasticity, with 77 clusters in LPMa1 and LPMa2, and 67 in LPMb1; standard errors in italics.

*** $p < 0.01$.

** $p < 0.05$.

* $p < 0.1$.

Table C.3b
Linear Restriction Tests on models presented in Table C.3a.

Null Hypothesis	*P* value on two-tailed test			
	LPMa1	**LPMb1**	**Metareg1**	**LPMa2**
Identity				
Ethnicity = Religion	0.613	0.652	0.872	0.43
Ethnicity = Nationality	0.553	0.157	0.991	0.256
Ethnicity = Gender	0.705	0.452	0.001***	< 0.001***
Ethnicity = Soc/GeoGroupings	< 0.001***	< 0.001***	< 0.001***	0.066*
Ethnicity = Natural Other	0.203	0.983	0.642	0.164
Religion = Nationality	0.909	0.571	0.884	0.126
Religion = Gender	0.496	0.431	0.009***	< 0.001***
Religion = Soc/Geo Groupings	0.001***	0.002***	< 0.001***	0.544
Religion = Natural Other	0.47	0.735	0.586	0.902
Nationality = Gender	0.286	0.017**	0.001***	0.085*
Nationality = Soc/Geo Groupings	< 0.001***	0.007***	< 0.001***	0.017**
Nationality = Natural Other	0.374	0.424	0.653	0.017**
Gender = Soc/Geo Groupings	< 0.001***	< 0.001***	< 0.001***	< 0.001***
Gender = Natural Other	0.169	0.769	0.16	< 0.001***
Soc/Geo Groupings = Natural Other	0.092*	0.004***	0.003***	0.452
Role Types				
TG Sender = TG Returner	0.25	0.171	0.119	0.53
TG Sender = Social Dilemma	0.605	0.648	0.782	0.808
TG Sender = Allocator	< 0.001***	0.008***	< 0.001***	0.2
TG Sender = Proposer	0.709	0.343	0.887	0.030**
TG Sender = Responder	0.891	0.25	0.37	0.758
TG Sender = Game Other	0.381	0.282	0.748	0.907
TG Returner = Social Dilemma	0.206	0.176	0.31	0.526
TG Returner = Allocator	< 0.001***	0.001***	< 0.001***	0.138
TG Returner = Proposer	0.647	0.863	0.341	0.016**
TG Returner = Responder	0.414	0.061*	0.085*	0.551
TG Returner = Game Other	0.086*	0.050*	0.155	0.758
Social Dilemma = Allocator	0.014***	0.019**	< 0.001***	0.398
Social Dilemma = Proposer	0.415	0.213	0.932	0.058*
Social Dilemma = Responder	0.761	0.458	0.309	0.912
Social Dilemma = Game Other	0.788	0.564	0.616	0.726
Allocator = Proposer	< 0.001***	< 0.001***	< 0.001***	0.158
Allocator = Responder	0.005***	0.107	< 0.001***	0.626
Allocator = Game Other	0.018**	0.050**	< 0.001***	0.234
Proposer = Responder	0.722	0.093*	0.367	0.183
Proposer = Game Other	0.186	0.021**	0.699	0.034**
Responder = Game Other	0.542	0.806	0.554	0.699

Note: LPMa is linear probability model run on full sample, Metareg is meta-regression run on sample for which effect sizes are available, LPMb is linear probability model run on same sample as Metareg.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

Table C.4
Linear probability regressions on discrimination and meta-regressions on effect size for artificial identity experiments only.

Dependent variable	Discrimination LPM	d Metareg
Role Types		
First Mover	−0.190	−0.074
	(0.165)	(0.127)
Second Mover	−0.102	−0.098
	(0.142)	(0.135)
Simultaneous Mover	−0.211	−0.005
	(0.154)	(0.139)
Allocator	0.236*	0.898***
	(0.130)	(0.170)
Partner Chooser	0.061	0.083
	(0.213)	(0.182)
Controls		
Students	0.032	0.128
	(0.123)	(0.207)
Sample Size	0.002***	0.001
	(0.001)	(0.001)
Identity Inducement Method		
Labelling	−0.117	0.030
	(0.081)	(0.085)
Other Method	−0.140	0.102
	(0.125)	(0.141)
Team Building	0.085	0.031
	(0.095)	(0.691)
Constant	0.409*	0.102
	(0.207)	(0.247)
R^2 (adjusted for Metareg)	0.154	0.262
N	169	146

Notes: LPM is linear probability model run on artificial identity sample, Metareg is meta-regression run on artificial identity sample for which effect sizes are available; omitted categories are Dictator (role type) and Preferences (Identity inducement method); errors in LPM are corrected for heteroskedasticity, with 32 clusters; standard errors in parentheses.
* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Table C.5
Linear probability regressions on discrimination and out-group favouritism, and meta-regression on effect size, with country-level variables included.

Dependent variable	Discrimination	d	Out-group favouritism		
	LPMa1	Metareg1	LPMa2	LPMa3	LPMa4
Identity					
Ethnicity	−0.220**	−0.171	0.053	0.044	0.042
	(0.086)	(0.104)	(0.045)	(0.045)	(0.042)
Religion	−0.146	−0.198	0.014	−0.063	−0.069
	(0.192)	(0.179)	(0.040)	(0.090)	(0.079)
Gender	−0.294***	−0.386***	0.254***	0.244***	0.239***
	(0.100)	(0.113)	(0.047)	(0.048)	(0.045)
Soc/Geo Groupings	0.260**	0.273***	−0.029*	−0.052*	−0.055**
	(0.103)	(0.095)	(0.016)	(0.029)	(0.025)
Natural Other	0.008	−0.311	−0.023	−0.093	−0.099*
	(0.177)	(0.213)	(0.017)	(0.068)	(0.057)
Role Types					
First Mover	−0.027	0.245***	−0.013	−0.012	−0.012
	(0.101)	(0.088)	(0.024)	(0.025)	(0.025)
Second Mover	−0.086	0.182*	0.014	0.014	0.014
	(0.080)	(0.094)	(0.025)	(0.026)	(0.026)
Simultaneous	−0.080	0.206*	−0.004	−0.004	−0.006
	(0.108)	(0.110)	(0.027)	(0.028)	(0.029)
Allocator	0.423***	1.178***	−0.026	−0.046	−0.049*
	(0.142)	(0.165)	(0.020)	(0.029)	(0.027)
Partner Chooser	0.058	0.208	0.049	0.067	–0.066
	(0.130)	(0.141)	(0.060)	(0.063)	(0.063)
Controls					
Fractionalisation	0.015	0.414			
	(0.265)	(0.257)			
LogGDPpc	0.020	−0.031			
	(0.083)	(0.089)			
Europe	0.100	0.277*			
	(0.131)	(0.150)			
USA	0.038	0.096	−0.049**	−0.012	
	(0.132)	(0.163)	(0.022)	(0.034)	
Individualism	−0.001	−0.001		−0.002	−0.002*
	(0.004)	(0.003)		(0.002)	(0.001)
Constant	0.191	0.174	0.035*	0.167	0.179*
	(0.731)	(0.813)	(0.021)	(0.119)	(0.094)
***R*² (adjusted in Metareg1)**	0.217	0.256	0.112	0.122	0.122
N	345	304	359	345	345

Notes: LPMa1, LPMa3 and LPMa4 are linear probability models run on full sample excluding experiments conducted across countries and in countries for which data on Individualism is not available, LPMa2 is linear probability model run on full sample excluding experiments conducted across countries, Metareg1 is meta-regression run on sample for which effect sizes are available excluding experiments conducted across countries and in countries for which data on Individualism is not available; omitted categories are Dictator (role type) and Artificial (identity); errors in LPM models are corrected for heteroskedasticity, with 60 clusters in LPMa1, LPMa3 and LPMa4, and 65 in LPMa2; standard errors in parentheses.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

Table C.6a
Linear probability regressions on discrimination and meta-regressions on effect size, with scope only for taste-based discrimination (Model 1) and scope for both types of discrimination (Model 2).

Dependent variable	Discrimination			d		
	Taste-based only LPMa1	Taste + Statistical LPMa2	Test of coefficient difference	Taste-based only Metareg1	Taste + Statistical Metareg2	Test of coefficient difference
Identity						
Ethnicity	−0.351***	−0.195*		−0.291***	0.020	
	(0.078)	(0.100)		(0.088)	(0.169)	
Religion	−0.474***	−0.159		−0.443**	−0.048	
	(0.125)	(0.234)		(0.172)	(0.206)	
Nationality	−0.292***	−0.162		−0.312***	−-0.133	
	(0.089)	(0.122)		(0.115)	(0.113)	
Gender	−0.290***	−0.352***		−0.474***	−0.642***	
	(0.090)	(0.080)		(0.122)	(0.167)	
Soc/Geo Groupings	0.081	0.408***	**	0.015	0.532***	***
	(0.159)	(0.126)		(0.125)	(0.142)	
Natural Other	0.146	−0.224		−0.301	−0.227	
	(0.263)	(0.262)		(0.329)	(0.281)	
Controls						
Students	0.041	−0.047		0.034	0.178	
	(0.101)	(0.095)		(0.101)	(0.142)	
Sample Size	$6.6e^{-4}$	$4.9e^{-4}$		$1.7e^{-5}$	$-3.1e^{-4}$	
	$(4.8e^{-4})$	$(7.2e^{-4})$		$(5.4e^{-4})$	$(7.2e^{-4})$	
Constant	0.380***	0.383***		0.355***	0.178	
	(0.117)	(0.120)		(0.114)	(0.153)	
R^2 **(adjusted in Metaregs)**	0.175	0.196		0.117	0.174	
N	262	179		204	160	

Notes: LPMa1 is linear probability model run on the sample for which discrimination can only be taste-based, LPMa2 is linear probability model run on the sample for which dicrimination can be both taste-based and statistical, Metareg1 is meta-regression run on the sample for which discrimination can only be taste-based and effect sizes are available, Metareg2 is meta-regression run on the sample for which discrimination can be both taste-based and statistical and effect sizes are available; 'test of coefficient difference' reports whether coefficients differ significantly between models 1 and 2; the omitted category is Artificial (identity); errors in LPM models are corrected for heteroskedasticity, with 65 clusters in LPMa1 and 59 in LPMa2; standard errors in parentheses.

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Table C.6b
Linear Restriction Tests on models presented in Table C.6a.

Null Hypothesis	***P* value on two-tailed test**			
	LPMa1	**LPMa2**	**Metareg1**	**Metareg2**
Ethnicity = Religion	0.171	0.876	0.38	0.756
Ethnicity = Nationality	0.353	0.788	0.87	0.405
Ethnicity = Gender	0.348	0.045**	0.163	0.003***
Ethnicity = Soc/Geo Groupings	0.002***	< 0.001***	0.016**	0.005***
Ethnicity = Natural Other	0.052*	0.911	0.977	0.373
Religion = Nationality	0.082*	0.989	0.509	0.699
Religion = Gender	0.097*	0.413	0.878	0.020**
Religion = Soc/Geo Groupings	< 0.001***	0.013**	< 0.005***	0.006***
Religion = Natural Other	0.018**	0.839	0.679	0.532
Nationality = Gender	0.985	0.057*	0.29	0.005***
Nationality = Soc/Geo Groupings	0.012**	< 0.001***	0.04**	< 0.001***
Nationality = Natural Other	0.099*	0.823	0.974	0.745
Gender = Soc/Geo Groupings	0.015**	< 0.001***	0.003***	< 0.001***
Gender = Natural Other	0.097*	0.624	0.613	0.192
Soc/Geo Groupings = Natural Other	0.81	0.015**	0.33	0.007

* $p < 0.1$.
** $p < 0.05$.
*** $p < 0.01$.

Appendix D: Papers included in the meta-analysis

Abbink, K., Harris, D., 2012. In-group favouritism and out-group discrimination in naturally occurring groups, Technical report, Mimeo, Monash University.
Ahmed, A.M., 2010. What is in a surname? The role of ethnicity in economic decision making. Appl. Econ. 42 (21), 2715-2723.
Ahmed, A.M., 2007. Group identity, social distance and intergroup bias. J. Econ. Psychol. 28 (3), 324-337.
Banuri, S., et al., 2012. Deconstructing nepotism. Available at: SSRN 2248187.
Bauernschuster, S., et al., 2009. Social Identity, Competition, and Finance: A Laboratory Experiment, Jena economic research papers.
Ben-Ner, A., et al., 2004. Share and share alike? Gender-pairing, personality, and cognitive ability as determinants of giving. J. Econ. Psychol. 25 (5), 581-589.
Bernhard, H., et al., 2006. Group affiliation and altruistic norm enforcement. Am. Econ. Rev. 96 (2), 217-221.
Binzel, C., Fehr, D., 2013. Social distance and trust: experimental evidence from a slum in Cairo. J. Dev. Econ. 103, 99-106.
Boarini, R., et al., 2009. Interpersonal comparisons of utility in bargaining: evidence from a transcontinental ultimatum game. Theory Decis. 67 (4), 341-373.
Bouckaert, J., Dhaene, G., 2004. Inter-ethnic trust and reciprocity: results of an experiment with small businessmen. Eur. J. Political Econ. 20 (4), 869-886.
Brandts, J., Sola, C., 2010. Personal relations and their effect on behavior in an organizational setting: an experimental study. J. Econ. Behav. Organ. 73 (2), 246-253.
Buchan, N.R., et al., 2006. Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preferences. J. Econ. Behav. Organ. 60 (3), 373-398.
Büchner, S., Dittrich, D.A., 2002. I will survive! - Gender discrimination in a household saving decisions experiment, Max Planck Institute of Economics, Strategic Interaction Group.
Burns, J., 2004. Race and trust in post-Apartheid South Africa. University of Cape Town, Centre for Social Science Research working paper 78.
Butler, J.V., 2014. Trust, truth, status and identity: an experimental inquiry. BE J. Theor. Econ. 14 (1), 293–338.
Carpenter, J., Cardenas, J.C., 2011. An intercultural examination of cooperation in the commons. J. Confl. Resolut. 55 (4), 632-651.
Chakravarty, S., Fonseca, M.A., 2013. Discrimination via Exclusion: An Experiment on Group Identity and Club Goods. University of Exeter Economics Department Discussion Papers Series. 13/02.
Charness, G., et al., 2007. Individual behavior and group membership. Am. Econ. Rev., 1340-1352.
Chen, R., Chen, Y., 2011. The potential of social identity for equilibrium selection. Am. Econ. Rev. 101 (6), 2562-2589.
Chen, Y., Li, S.X., 2009. Group identity and social preferences. Am. Econ. Rev. 99 (1), 431-457.
Chen, Y., et al., 2014. Which hat to wear? Impact of natural identities on coordination and cooperation. Games Econ. Behav. 84, 58-86.
Chuah, S.H., et al., 2007. Do cultures clash? Evidence from cross-national ultimatum game, experiments. J. Econ. Behav. Organ. 64, 35-48.
Chuah, S.-H., et al., 2013. Fractionalization and trust in India: a field-experiment. Econ. Lett. 119 (2), 191-194.
Costard, J., Bolle, F., 2011. Solidarity, responsibility and group identity, Discussion paper, European University Viadrina, Department of Business Administration and Economics.
Currarini, S., Mengel, F., 2012. Identity, Homophily and In-Group Bias. 〈services.bepress.com〉.
Daskalova, V., 2012. Discrimination, Social Identity, and Coordination: An Experiment. 〈webspace.qmul.ac.uk〉.
Delavande, A., Zafar, B., 2011. Stereotypes and Madrassas. Federal Bank of New York Staff Reports 501.
Der Merwe, V., et al., 2008. What's in a name? Racial identity and altruism in post apartheid South Africa. South Afr. J. Econ. 76 (2), 266-275.
Dugar, S., Shahriar, Q., 2010. Group Identity and the Moral Hazard Problem: Evidence From the field. San Diego State University, Department of Economics Working Papers.
Eckel, C.C., Grossman, P.J., 2001. Chivalry and solidarity in ultimatum games. Econ. Inq. 39 (2), 171-188.
Etang, A., et al., 2011a. Does trust extend beyond the village? Experimental trust and social distance in Cameroon. Exp. Econ. 14 (1), 15-35.
Etang, A., et al., 2011b. Trust and rosca membership in rural cameroon. J. Int. Dev. 23 (4), 461-475.
Fehr, E., et al., 2013. The development of egalitarianism, altruism, spite and parochialism in childhood and adolescence. Eur. Econ. Rev. 64, 369-383.
Ferraro, P.J., Cummings, R.G., 2007. Cultural diversity, discrimination, and economic outcomes: an experimental analysis. Econ. Inq. 45 (2), 217-232.
Fershtman, C., et al., 2005. Discrimination and nepotism: the efficiency of the anonymity rule. J. Leg. Stud. 34 (2), 371-394.
Fiedler, M., et al., 2011. Social distance in a virtual world experiment. Games Econ. Behav. 72 (2), 400-426.
Filippin, A., Guala, F., 2013. Costless discrimination and unequal achievements in an experimental tournament. Exp. Econ. 16 (3), 285-305.
Finocchiaro Castro, M., 2008. Where are you from? Cultural differences in public good experiments. J. Socio-Econ. 37 (6),

2319-2329.
Fong, C.M., Luttmer, E.F., 2011. Do fairness and race matter in generosity? Evidence from a nationally representative charity experiment. J. Public Econ. 95 (5), 372-394.
Friesen, J., et al., 2012. Ethnic identity and discrimination among children. J. Econ. Psychol. 33 (6), 1156-1169.
Georg, S., et al., 2008. Distributive Fairness in an Intercultural Ultimatum Game, Jena economic research papers.
Gneezy, U., et al., 2012. Toward an Understanding of Why People Discriminate: Evidence from a Series of Natural Field Experiments. National Bureau of Economic Research, New York.
Goette, L., et al., 2006. The impact of group membership on cooperation and norm enforcement: evidence using random assignment to real social groups. Am. Econ. Rev. 96 (2), 212-216.
Goette, L., et al., 2012a. The impact of social ties on group interactions: evidence from minimal groups and randomly assigned real groups. Am. Econ. J. - Microecon. 4 (1), 101-115.
Goette, L., et al., 2012b. Competition between organizational groups: its impact on altruistic and antisocial motivations. Manag. Sci. 58 (5), 948-960.
Guala, F., et al., 2013. Group membership, team preferences, and expectations. J. Econ. Behav. Organ. 86, 183-190.
Guillen, P., Ji, D., 2011. Trust, discrimination and acculturation: experimental evidence on Asian international and Australian domestic university students. J. Socio-Econ. 40 (5), 594-608.
Gupta, G., et al., 2013. Religion, Minority Status and Trust: Evidence from a Field Experiment. Monash University, Department of Economics, Australia.
Guth, W., et al., 2005. The Effect of Group Identity in an Investment Game. Papers on Strategic Interaction.
Guth, W., et al., 2009. Determinants of in-group bias: Is group affiliation mediated by guilt-aversion? J. Econ. Psychol. 30 (5), 814-827.
Haile, D., et al., 2008. Cross-racial envy and underinvestment in South African partnerships. Camb. J. Econ. 32 (5), 703-724.
Hargreaves Heap, S.P., Zizzo, D.J., 2009. The value of groups. Am. Econ. Rev. 99 (1), 295-323.
Harris, D., et al., 2009. Two's Company, Three's a Group: The Impact of Group Identity and Group Size on In-Group Favouritism, CeDEx discussion paper series.
Hennig-Schmidt, H., et al., 2007. Actions and Beliefs in a Trilateral Trust Game Involving Germans, Israelis and Palestinians (unpublished manuscript).
Hoff, K., et al., 2011. Caste and punishment: the legacy of caste culture in norm enforcement. Econ. J. 121 (556), F449-F475.
Hopfensitz, A., Miquel-Florensa, P., 2013. Public Good Contributions Among Coffee Farmers in Costa Rica: Cooperativists and Private Market Participants. Toulouse School of Economics (TSE), France.
Houser, D., Schunk, D., 2009. Social environments with competitive pressure: Gender effects in the decisions of German schoolchildren. J. Econ. Psychol. 30 (4), 634-641.
Ioannou, C.A., et al., 2013. Group Payoffs As Public Signals. ⟨christosaioannou.com⟩ (accessed 10.12.13).
Johansson-Stenman, O., et al., 2009. Trust and religion: experimental evidence from rural Bangladesh. Economica 76 (303), 462-485.
Kim, B.-Y., et al., 2013. Do institutions affect social preferences? Evidence from divided Korea, IZA Discussion Paper.
Lankau, M., et al., 2012. Cooperation preferences in the provision of public goods: an experimental study on the effects of social identity, Discussion Papers, Center for European Governance and Economic Development Research.
Li, S.X., et al., 2011. Group identity in markets. Int. J. Ind. Organ. 29 (1), 104-115.
Masella, P., et al., 2014. Incentives and group identity. Games Econ. Behav. 86, 12-25.
McLeish, K.N., Oxoby, R.J., 2007. Identity, Cooperation, and Punishment, IZA Discussion Papers, No. 2572.
Morita, H., Servátka, M., 2013. Group identity and relation-specific investment: an experimental investigation. Eur. Econ. Rev. 58, 95-109.
Netzer, R.J., Sutter, M., 2009. Intercultural trust. An experiment in Austria and Japan, Working Papers in Economics and Statistics.
Ortmann, A., Tichy, L.K., 1999. Gender differences in the laboratory: evidence from prisoner's dilemma games. J. Econ. Behav. Organ. 39 (3), 327-339.
Pecenka, C.J., Kundhlande, G., 2013. Theft in South Africa: an experiment to examine the influence of racial identity and inequality. J. Dev. Stud. 49 (5), 737-753.
Ploner, M., Soraperra, I., 2004. Groups and Social Norms in the Economic Context: A Preliminary Experimental Investigation, Cognitive and Experimental Economics Laboratory. Department of Economics, University of Trento, Italia.
Ruffle, B.J., Sosis, R., 2006. Cooperation and the in-group-out-group bias: a field test on Israeli kibbutz members and city residents. J. Econ. Behav. Organ. 60 (2), 147-163.
Shahriar, Q., 2011. Identity in a second-price sealed bid auction: an experimental investigation. Manch. Sch. 79 (1), 159-170.
Slonim, R., Guillen, P., 2010. Gender selection discrimination: evidence from a trust game. J. Econ. Behav. Organ. 76 (2), 385-405.
Solnick, S.J., 2001. Gender differences in the ultimatum game. Econ. Inq. 39 (2), 189-200.
Spiegelman, E., 2012. «C'EST CE QUE JE VOUS DIS»: ESSAIS SUR L'ANALYSE ÉCONOMIQUE DE LA COMMUNICATION INTERPERSONNELLE.
Tremewan, J., 2010. Group Identity and Coalition Formation: Experiments in One-Shot and Repeated Games. ⟨webmeets.com⟩.

Walkowitz, G., et al., 2004. Experimenting over a Long Distance: A Method to Facilitate Intercultural Experiments, Bonn econ discussion papers.
Willinger, M., et al., 2003. A comparison of trust and reciprocity between France and Germany: experimental investigation based on the investment game. J. Econ. Psychol. 24 (4), 447-466.
Wu, F., 2009. Cultural Affinity in International Joint Ventures-An Experimental Study. 〈eale.nl conference paper〉.
Zizzo, D.J., 2011. You are not in my boat: common fate and discrimination against outgroup members. Int. Rev. Econ. 58 (1), 91-103.

Appendix E. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.euroecorev.2015.11.011.

References

Other sources cited:

Aberson, C.L., et al., 2000. Ingroup bias and self-esteem: a meta-analysis. Personal. Soc. Psychol. Rev. 4 (2), 157–173.
Alesina, A., et al., 2003. Fractionalization. J. Econ. Growth 8 (2), 155–194.
Anderson, J., et al., 2013. Self-selection and variations in the laboratory measurement of other-regarding preferences across subject pools: evidence from one college student and two adult samples. Exp. Econ. 16 (2), 170–189.
Arrow, K., 1972. Some mathematical models of race discrimination in the labor market. In: Pascal, A.H. (Ed.), Racial Discrimination in Economic Life, D.C. Heath, Lexington, MA, pp. 187–204.
Balliet, D., et al., 2014. Ingroup favoritism in cooperation: a meta-analysis. Psychol. Bull. 140 (6), 1556–1581.
Becker, G.S., 2010. In: The Economics of DiscriminationUniversity of Chicago press, United States.
Bellemare, C., Kröger, S., 2007. On representative social capital. Eur. Econ. Rev. 51 (1), 183–202.
Bettencourt, B., et al., 2001. Status differences and in-group bias: a meta-analytic examination of the effects of status stability, status legitimacy, and group permeability. Psychol. Bull. 127 (4), 520.
Blanchard, F.A., et al., 1994. Condemning and condoning racism: a social context approach to interracial settings. J. Appl. Psychol. 79 (6), 993.
Buchan, N.R., et al., 2009. Globalization and human cooperation. Proc. Natl. Acad. Sci. USA 106 (11), 4138–4142.
Crandall, C.S., et al., 2002. Social norms and the expression and suppression of prejudice: the struggle for internalization. J. Personal. Soc. Psychol. 82 (3), 359.
Engel, C., 2007. How much collusion? A meta-analysis of oligopoly experiments. J. Compét. Law Econ. 3 (4), 491–549.
Engel, C., 2011. Dictator games: a meta study. Exp. Econ. 14 (4), 583–610.
Falk, A., Zehnder, C., 2007. Discrimination and in-group favoritism in a citywide trust experiment, IZA Discussion Papers.
Fischer, R.D., Derham, C., 2010. Is the Minimal Group Paradigm Culture dependent? A Cross-Cultural Multi-level Analysis. 〈http://www.psychology.org.au/ext/iaccp2010/saturday-10-july/6/0830/fischer-r.pdf〉 (accessed 05.11.13).
Harbord, R.M., Higgins, J.P., Meta-Regression in Stata. Meta 8.4, 2008, pp. 493–519.
Hedges, L.V., Olkin, I., 1985. Statistical Methods for Meta-Analysis. Academic Press, Orlando, FL.
Hopfensitz, A., 2009. Previous Outcomes and Reference Dependence: A Meta Study of Repeated Investment Tasks with and Without Restricted Feedback. 〈mpra.ub.uni-muenchen.de〉 (accessed 10.11.13).
Johnson, N.D., Mislin, A.A., 2011. Trust games: a meta-analysis. J. Econ. Psychol. 32 (5), 865–889.
Jones, G., 2008. Are smarter groups more cooperative? Evidence from prisoner's dilemma experiments, 1959–2003. J. Econ. Behav. Organ. 68 (3), 489–497.
Kiyonari, T., Yamagishi, T., 2004. In-group cooperation and the social exchange heuristic. In: Suleiman, R., Budescu, D.V., Fischer, I., Messick, D.M. (Eds.), Contemporary Psychological Research on Social Dilemmas, Cambridge University Press, New York, NY, pp. 269–286.
Oxford Dictionaries, definition of ‘political correctness’. 〈http://www.oxforddictionaries.com/definition/english/political-correctness〉 (accessed 01.03.15).
Percoco, M., Nijkamp, P., 2009. Estimating individual rates of discount: a meta-analysis. Appl. Econ. Lett. 16 (12), 1235–1239.
Prante, T., et al., 2007. Evaluating coasean bargaining experiments with meta-analysis. Econ. Bull. 3 (68), 1–7.
Robbins, J.M., Krueger, J.I., 2005. Social projection to ingroups and outgroups: a review and meta-analysis. Personal. Soc. Psychol. Rev. 9 (1), 32–47.
Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychol. Bull. 86 (3), 638.
Rothstein, H.R., et al., 2006. Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments. John Wiley & Sons, United States.
Saucier, D.A., et al., 2005. Differences in helping whites and blacks: a meta-analysis. Pers. Soc. Psychol. Rev. 9 (1), 2–16.
Tajfel, H., et al., 1971. Social Categorization and Intergroup Behavior. Eur. J. Soc. Psychol. 1 (2), 149–177.
Weizsäcker, G., 2010. Do we follow others when we should? A simple test of rational expectations. Am. Econ. Rev. 100, 2340–2360.
Zizzo, D.J., 2010. Experimenter demand effects in economic experiments. Exp. Econ. 13 (1), 75–98.