CEDEX
CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Discussion Paper No. 2015-03

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

Tom Lane

March 2015

CeDEx Discussion Paper Series
ISSN 1749 - 3293

CEDEX

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Suzanne Robey
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 14763
Fax: +44 (0) 115 95 14159
suzanne.robey@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

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

Tom Lane, University of Nottingham

March 16, 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 447 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 seem to drive the relatively strong discrimination with artificial identity, while statistical motivations moderate it. 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 (1957), and with the growth of experimental economics in the last two decades, experiments have emerged as a popular complement to survey-based econometric studies.

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. They can also investigate quite specific behavioural determinants of discrimination, in particular variations in trust levels – indeed, trust games have become some of the most popular experimental tools in this field. Furthermore, through the use of incentives and between-subject designs, 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.

[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; Weiszäcker (2010) – social learning; Engel (2011) – dictator games; Johnson and Mislin (2011) – trust games.

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 447 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.

A meta-analysis can help uncover underlying patterns of behaviour in these experiments, and inform the designers of future ones. For example, it can reveal how the strength of discrimination tends to vary according to the type of identity being investigated: artificial (i.e. using minimal groups) versus such natural types as ethnicity, nationality, religion, gender and social/geographical affiliation. Additionally, it can shed light on whether and how discrimination varies by game setting, or by whether the sample pool is composed of students or non-students.

A meta-analysis can also shed light on what motivates discrimination and how this varies by context. 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, with little consensus on which is more prescient in many cases. The taste-based model, proposed by Becker (1957), entails individuals gaining direct utility from the act of discriminating against out-groups. Meanwhile, statistical theories – beginning with Arrow (1972) – posit discrimination is based on beliefs about average group differences which, in the presence of imperfect information, lead economic actors to optimise their actions conditional on the group membership of those they interact with.

In this paper, we ask the following questions. (1) On average, what is the extent of discrimination in this literature? (2) Does the level of discrimination vary according to the type of identity groups are based upon? (3) 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 support for taste-based or statistical theories of discrimination, and (6) does the experimental context make either type of discrimination more or less important? (7) In gender experiments, does male-to-female discrimination differ from 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. Furthermore, (6) discrimination against the out-group emerges particularly strongly in artificial group experiments when only taste-based discrimination is possible, but it weakens when there is also the possibility of statistical discrimination; we interpret this as evidence that the relatively strong discrimination in artificial group experiments is driven by tastes and tempered by statistical beliefs. (7) 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.

We argue, given the above differences, that our analysis provides a more focused and accurate picture of the experimental economics literature on discrimination than that of Balliet et al. 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

[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 and Derham (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.

subcategories as we do. The clearest difference in results between the two studies is that Balliet et al find discrimination is stronger by trust game senders than by dictators, and stronger still in social dilemmas, 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, 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, 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 layer of mitigation against

[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.

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. 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 368 of our 447 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. We then do the same 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 for the two phenomena. 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).

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

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 those of Asian descent comprise a mixture of ethnicities.[5] Papers such as Falk and Zender (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

[5] There were four cases where we made such subjective decisions, all listed in Appendix 1. Our main results still hold regardless of the decisions we come to in these cases.

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.

Our inclusion criteria do not specifically create any distinction between lab and field experiments, but it turns out that no pure field experiment – in which subjects do not know they are participants in a study – meet all the criteria. Our meta-analysis does contain lab-in-the-field studies. We 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). While these studies reveal a lot about discrimination, their methodology differs from the experiments included in our analysis in, arguably, three respects: they do not (or cannot) report the identity of the respondents, and therefore cannot determine exactly which candidate is an out-group member to which decision-maker; they use deception; and real material payoffs are not necessarily at stake.

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 3, Table A2. 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 dealing with the continuous dependent variable, we use standard random effects meta-analysis procedures to determine average effect sizes. These are simply aggregate estimates – not taking into account any moderating factors – that first weight each result by the inverse of its standard error, thus attaching more importance to larger and more precise results, then follow an unweighting process, the extent of which depends upon the heterogeneity in effect sizes (Harbord and Higgins, 2008).[6]

We also apply random effects meta-regressions, which follow these same weighting processes and allow the inclusion of independent variables in the analysis. 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 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

Role type dummies: We include role type dummy variables to pursue the question of how different decision-making contexts affect the extent of discrimination. In some games, all participants face the same decision-making problem, while in others roles are asymmetric and should be treated separately. For example, trust games generate data on the behaviour of senders and returners, who are placed in different situations.

We gain the greatest number of observations from trust games and similar principal-agent games, which provide two roles: senders (*TG Sender*, 98 observations) and returners (*TG Returner*, 78). The next most common role type is the *Dictator* (66). 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). 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. Ultimatum games and similar bargaining settings are grouped together and split into two role types: first movers (*Proposer*, 31) and second movers (*Responder*, 29). 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* (51 observations) into which are placed the remaining game settings that we did not think could be adequately categorised.[7]

[6] This 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.

[7] 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.

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.[8] 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 175 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, 2009), colleges within universities (Banuri et al, 2012) or friendship groups (Brands 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 Miguel-Florensa, 2013). We pool them under the composite variable *Natural Other*[9].[10]

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

[8] 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 2, 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.

[9] 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.

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

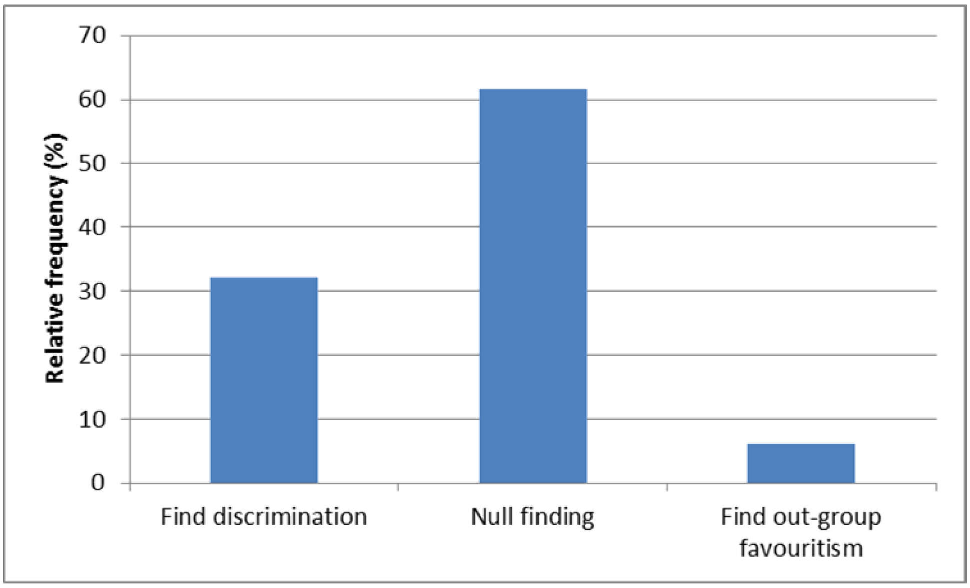
non-student participants. Likewise, while we accept experiments in the field may include a few student subjects, their proportion is likely to be low (unless otherwise stated). As another control, we include the size of the active decision-making sample from which a given result is derived (*Sample Size)*.

3. Results

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

In total, as shown in Figure 1, there are 144 results indicating significant discrimination (32.21%), 28 indicating significant out-group favouritism (6.18%), and 275 indicating no significant discrimination or out-group favouritism (61.52%). 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.

Figure 1: Breakdown of data-points by result type



For the sub-sample where we are able to generate effect sizes (364 of 447 observations), the random effects meta-analysis finds an overall effect size of 0.252 (95% confidence range: 0.205 - 0.3). 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).

Result 2: The strength of discrimination depends upon the type of group identity under investigation.

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 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		175	40.6	57.1	2.3	150	0.353 (0.266 – 0.439)
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 – -0.053)
	Religious	21	14.3	80.9	4.8	21	0.034 (-0.062 – 0.131)
	Soc/Geo Groupings	57	64.9	35.1	0.0	51	0.551 (0.432 – 0.669)
	Natural Other	17	47.1	52.9	0.0	7	-0.036 (-0.158 – 0.086)

The first three columns of Table 2 report LPM estimations for discrimination against the out-group, and a meta-regression on the discrimination effect size, to test whether these identity-type variables yield significantly different levels of discrimination while controlling for other factors. In all models artificial identity is the benchmark category. LPMb1 is a linear probability model run on the reduced sample for which effect size calculation is possible. In Appendix 4, Table A6 presents the results of linear restriction tests run on the sets of dummy variables featuring in the models in Table 2.

In LPMa1 the coefficients on the ethnic, national and gender identity types are significantly negative (at the 1% level for *Ethnicity* and *Gender*; at the 5% level for *Nationality*), strongly indicating that discrimination is less likely to be observed when subject pools are split along these lines than on the basis of identities artificially created during experiments. According to the equivalent meta-regression (Metareg1), however, national identity experiments are not linked to significantly lower discrimination (i.e. less positive effect sizes) than artificial group experiments, and ethnic identity experiments only are at 10% level. For nationality, this appears to be due to the change in sample, as in LPMb1 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[11]. Like the LPMs, the meta-regression indicates weaker gender discrimination than between artificial groups, significant at the 1% level.

In both the linear probability models and meta-regression, the only identity category linked to stronger discrimination than the artificial type is *Soc/Geo Groupings*, with the difference significant at the 1% level in LPMa1 and Metareg1. The *Soc/Geo Groupings* coefficient is, in linear restriction tests, found always to be significantly higher than the other identity type dummies at the 1% level – although the difference with the coefficient on *Natural Other* is only significant at the 10% level according to a test on LPMa1. Equivalent tests show *Gender* to be associated with lower discrimination than *Nationality, Religion* and *Ethnicity* in Metareg1, although the difference with *Religion* is only weakly significant.

Gender is also significant in LPMa2, a linear probability regression with out-group favouritism as the dependent variable, presented on the far right of Table 2. 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. Additionally, 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 in LPMa2, 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.

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 2.

[11] In Table 3, we will later present a meta-regression with the number of role type dummies reduced from seven 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 5% level in that model.

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

Dependent variable	Discrimination		d	Out-group favouritism
	LPMa1	LPMb1	Metareg1	LPMa2
Role Types				
TG Sender	-0.017	-0.008	0.093	0.010
	(0.079)	(0.091)	(0.079)	(0.028)
TG Returner	-0.095	-0.106	-0.010	0.026
	(0.080)	(0.092)	(0.085)	(0.025)
Social Dilemma	0.045	0.053	0.085	-2.8e^{-4}
	(0.109)	(0.120)	(0.093)	(0.039)
Allocator	0.377***	0.434***	1.302***	-0.024
	(0.098)	(0.116)	(0.131)	(0.034)
Proposer	-0.053	-0.096	0.074	-0.077*
	(0.104)	(0.096)	(0.100)	(0.041)
Responder	-0.026	0.164	0.204	-0.006
	(0.107)	(0.106)	(0.127)	(0.053)
Game Other	0.049	0.084	0.057	0.022
	(0.101)	(0.125)	(0.093)	(0.036)
Identity				
Ethnicity	-0.283***	-0.286***	-0.127*	0.034
	(0.076)	(0.081)	(0.075)	(0.041)
Religion	-0.230	-0.244	-0.159	-0.025
	(0.139)	(0.149)	(0.119)	(0.056)
Nationality	-0.224**	-0.126	-0.095	0.109
	(0.085)	(0.105)	(0.072)	(0.067)
Gender	-0.292***	-0.317***	-0.442***	0.237***
	(0.064)	(0.068)	(0.091)	(0.043)
Soc/Geo Groupings	0.230***	0.215**	0.284***	-0.052**
	(0.085)	(0.099)	(0.083)	(0.022)
Natural Other	-0.070	-0.296	-0.250	-0.037
	(0.179)	(0.181)	(0.180)	(0.035)
Controls				
Students	-0.014	-0.057	0.026	-0.019
	(0.061)	(0.068)	(0.073)	(0.041)
Sample Size	2.2e^{-4}	1.8e^{-4}	-4.0e^{-4}	2.9e^{-4}
	(4.3e^{-4})	(4.5e^{-4})	(4.1e^{-4})	(3.1e^{-4})
Constant	0.392***	0.422***	0.194**	0.025
	(0.087)	(0.098)	(0.098)	(0.047)
R^2 (adjusted in Metareg1)	0.204	0.213	0.337	0.095
N	447	368	368	447

Note: *** p<0.01, ** p<0.05, * p<0.1; 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 parentheses.

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

Inspection of the coefficients on role type dummies in LPMa1 and Metareg1 (Table 2) 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 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 (1.302) 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 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, Table 2), we find no significant differences at all between any role type pair. [12]

Result 4: Discrimination does not significantly differ between students and 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. 31.2% of the observations for students return discrimination, while 6.7% find out-group favouritism and 62.1% 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.

In Appendix 2, we also show that the country where an experiment is run is not a significant predictor of the extent of the discrimination found.

Result 5: There is evidence for both taste-based and statistical discrimination.

For 215 (48.5%) 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 this move is made simultaneously with others, such as in prisoners' dilemmas. Discrimination by dictators and trust game returners, for example, can

[12] In the analysis of the effect of role type, of interest is Kiyonari and Yamagishi (2004), who advance the view that discrimination is stronger in games where players move simultaneously, rather than one after the other. This is supported by Balliet et al (2014) (although Balliet et al only consider the behaviour of first movers in sequential exchanges). To test it, we drop observations for roles where opponents are passive, and re-code the rest into three categories: simultaneous movers, first movers in sequential exchanges, and responders in sequential exchanges. We do not find support for this view; as Table A4 (Appendix 3) shows, we see no significant differences in discrimination between these types of setting, either when using the binary or continuous dependent variable.

only be taste-based, because opponents then have no control over the final outcome and beliefs about their type are therefore irrelevant.[13]

In Table 3, we run linear probability regressions on discrimination and out-group favouritism, 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. There is no significant difference in the likelihood of observing discrimination or out-group favouritism (LPMa1 and LPMa2), or in the predicted effect size (Metareg), when scope for statistical discrimination is added.

This would suggest statistical discrimination is not an important driver of behaviour in these experiments, but we probe further by analysing the results of individual experiments. Where there is scope for statistically-motivated discrimination, by design for 74.1% 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. In the 62 cases that it is possible to distinguish between discriminatory motives, the authors find significant statistical discrimination to occur in 15 cases (11 times against the out-group and four 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).

It seems, then, 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 moderate 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.

[13] 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).

Table 3: Linear probability regressions on discrimination and out-group favouritism, and meta-regression on effect size, with or without scope for statistical discrimination.[14]

Dependent variable	Discrimination	d	Out-group favouritism
	LPMa1	Metareg	LPMa2
Type of discrimination possible			
Taste + Statistical	-0.044	-0.013	-0.004
	(0.054)	(0.053)	(0.017)
Identity			
Ethnicity	-0.267***	-0.184**	0.037
	(0.060)	(0.081)	(0.034)
Religion	-0.258*	-0.208	-0.007
	(0.135)	(0.131)	(0.049)
Nationality	-0.224***	-0.195**	0.112
	(0.074)	(0.078)	(0.068)
Gender	-0.306***	-0.534***	0.231***
	(0.063)	(0.100)	(0.047)
Soc/Geo Groupings	0.253***	0.279***	-0.044*
	(0.096)	(0.093)	(0.023)
Natural Other	0.091	-0.266	-0.039
	(0.255)	(0.204)	(0.027)
Controls			
Students	0.049	0.119	-0.026
	(0.074)	(0.082)	(0.039)
Sample Size	$4.2e^{-4}$	$-3.0e^{-4}$	$2.0e^{-4}$
	($4.6e^{-4}$)	($4.4e^{-4}$)	($2.8e^{-4}$)
Constant	0.315***	0.247**	0.034
	(0.100)	(0.098)	(0.054)
R^2 (adjusted in Metareg)	0.154	0.129 (adjusted)	0.085
N	447	368	447

Note: *** p<0.01, ** p<0.05, * p<0.1; omitted categories are taste-based only (type of discrimination possible) and Artificial (identity); errors in LPM models are corrected for heteroskedasticity, with 77 clusters; standard errors in parentheses.

Result 6: The relatively strong discrimination in artificial group experiments is driven by taste-based behaviour and moderated by belief-based behaviour.

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

[14] 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.

taste-based and statistical discrimination. The results are presented in Table 4; LPMa1 and Metareg1 relate to the taste-based only sub-sample, while LPMa2 and Metareg2 relate to the both-types sub-sample. The results of linear restriction tests are presented in Appendix 4, Table A7.

Table 4: 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)

	Taste-based only		Taste + Statistical	
Dependent variable	Discrimination	d	Discrimination	d
	LPMa1	Metareg1	LPMa2	Metareg2
Identity				
Ethnicity	-0.384***	-0.231	-0.210***	-0.149
	(0.118)	(0.140)	(0.074)	(0.101)
Religion	-0.508***	-0.403**	-0.149	-0.125
	(0.141)	(0.194)	(0.220)	(0.183)
Nationality	-0.317***	-0.356***	-0.130	-0.081
	(0.094)	(0.123)	(0.115)	(0.102)
Gender	-0.300***	-0.510***	-0.320***	-0.587***
	(0.097)	(0.130)	(0.068)	(0.154)
Soc/Geo Groupings	0.081	0.077	0.371***	0.459***
	(0.179)	(0.155)	(0.104)	(0.123)
Natural Other	0.270	-0.260	-0.191	-0.331
	(0.268)	(0.353)	(0.164)	(0.254)
Controls				
Students	0.025	0.115	-0.077	0.036
	(0.121)	(0.128)	(0.071)	(0.115)
Sample Size	6.1e^{-4}	-1.9e^{-4}	5.0e^{-4}	-6.8e^{-6}
	(4.6e^{-4})	(5.9e^{-4})	(6.9e^{-4})	(6.6e^{-4})
Constant	0.423***	0.325**	0.380***	0.257**
	(0.135)	(0.138)	(0.103)	(0.127)
R^2 (adjusted in Metaregs)	0.194	0.131	0.170	0.145
N	215	161	232	207

Note: *** p<0.01, ** p<0.05, * p<0.1; the omitted category is Artificial (identity); errors in LPM models are corrected for heteroskedasticity, with 58 clusters in LPMa1 and 65 in LPMa2; standard errors in parentheses.

When it can only be driven by taste, according to the LPM discrimination is much greater across artificial groups than across ethnicities, religions, nationalities or gender – all differences significant at the 1% level – and not significantly different to across socially or geographically distinct groups; results are similar in the meta-regression, although the artificial and ethnicity effect sizes are not found to be significantly different. When there is scope for both types, however, discrimination is not in either model found to be significantly different between artificial group experiments and those on nationality or religion, and is significantly higher (at the 1% level) in conditions with socially or geographically distinct groups. In the LPM the coefficient on *Ethnicity* is also rather closer to zero when there is scope for both types of discrimination. We interpret this as tastes driving a bias for discrimination in artificial group experiments – relative to national, religious and perhaps ethnic identity experiments – and beliefs moderating it. In all models linear restriction tests show discrimination between social or geographical groupings to be significantly stronger than between groups of other types of natural identity (apart from the *Natural Other* category), so we interpret a change in the strength of the artificial identity effect to be the reason why *Soc/Geo Groupings* is insignificant when there is only scope for taste-based discrimination and very significantly positive when there is scope for both types.

Result 7: As noted above, there is significant out-group favouritism in gender experiments. Females significantly favour males; males favour females but the effect is only weakly significant.

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 – -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. All three results of discrimination against the other gender are for female decision-makers, while six of the eight results of other-gender favouritism are for male decision-makers. 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 - -0.013) for females and -0.173 (95% confidence range: -0.369 – 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.

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'[15] 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, 2003), 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). Our analysis finds the strong discrimination in artificial group experiments, relative to ethnic and national identity experiments, to be driven by tastes and tempered by statistical beliefs. Perhaps ethnic or national discrimination on the basis of beliefs feels

[15] 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).

more socially acceptable, or at least less blatant, than discrimination on the basis of a raw preference for one type of person over another; or perhaps focusing attention on what one's counterpart might do distracts from thinking about social acceptability. 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[16].

If it were shown that discrimination in economics experiments is indeed mitigated 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 (2012) 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 an intrinsic aspect of laboratory-dependent natural group bias as to in fact often go beyond it. On the other hand, if experimenters' goal is to learn how discrimination differs according to aspects of the decision-making context besides the precise relationship between two given groups, they would be best served using artificial groups, as this generates a relatively good chance of finding significant discrimination in at least one treatment – and any findings cannot be attributed to the idiosyncratic behaviour of the natural groups in the experiment.

[16] 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.

Regarding role type, we find discrimination by third-party allocators is much stronger than by participants in any other game setting. If social acceptability is indeed a moderator of 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 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, the effect of this role presents experimenters with a dilemma: they are more likely to identify significant discrimination if they employ the role, but should therefore be less confident about the out-of-context generalizability of such results.

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.

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 Network for Integrated Behavioural Science for financial support.

References

Papers included in the meta-analysis:

Abbink, K. and D. Harris (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." Applied Economics 42.21: 2715-2723.

Ahmed, A. M. (2007). "Group identity, social distance and intergroup bias." Journal of Economic Psychology 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." Journal of Economic Psychology 25(5): 581-589.

Bernhard, H., et al. (2006). "Group affiliation and altruistic norm enforcement." American Economic Review 96(2): 217-221.

Binzel, C. and D. Fehr (2013). "Social distance and trust: Experimental evidence from a slum in Cairo." Journal of Development Economics 103: 99-106.

Boarini, R., et al. (2009). "Interpersonal comparisons of utility in bargaining: evidence from a transcontinental ultimatum game." Theory and decision 67(4): 341-373.

Bouckaert, J. and G. Dhaene (2004). "Inter-ethnic trust and reciprocity: results of an experiment with small businessmen." European Journal of Political Economy 20(4): 869-886.

Brandts, J. and C. Sola (2010). "Personal relations and their effect on behavior in an organizational setting: An experimental study." Journal of Economic Behavior & Organization 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." Journal of Economic Behavior & Organization 60(3): 373-398.

Büchner, S. and D. A. Dittrich (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. The BE Journal of Theoretical Economics, 14(1).

Carpenter, J. and J. C. Cardenas (2011). "An Intercultural Examination of Cooperation in the Commons." Journal of Conflict Resolution 55(4): 632-651.

Chakravarty, S. and M. A. Fonseca (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." The American Economic Review: 1340-1352.

Chen, R. and Y. Chen (2011). "The potential of social identity for equilibrium selection." The American Economic Review 101(6): 2562-2589.

Chen, Y. and S. X. Li (2009). "Group Identity and Social Preferences." American Economic Review 99(1): 431-457.

Chen, Y., et al (2014). Which hat to wear? Impact of natural identities on coordination and cooperation. Games and Economic Behavior, 84, 58-86.

Chuah, S. H., et al. (2007). "Do cultures clash? Evidence from cross-national ultimatum game, experiments." Journal of Economic Behavior & Organization 64(1): 35-48.

Chuah, S.-H., et al. (2013). "Fractionalization and trust in India: A field-experiment." Economics Letters 119(2): 191-194.

Costard, J. and F. Bolle (2011). Solidarity, responsibility and group identity, Discussion paper//European University Viadrina, Department of Business Administration and Economics.

Currarini, S. and F. Menge (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. and B. Zafar (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 African Journal of Economics 76(2): 266-275.

Dugar, S. and Q. Shahriar (2010). "Group identity and the moral hazard problem: Evidence from the field." San Diego State University, Department of Economics Working Papers.

Eckel, C. C. and P. J. Grossman (2001). "Chivalry and solidarity in ultimatum games." Economic Inquiry 39(2): 171-188.

Etang, A., et al. (2011). "Does trust extend beyond the village? Experimental trust and social distance in Cameroon." Experimental economics 14(1): 15-35.

Etang, A., et al. (2011). "Trust and rosca membership in rural cameroon." Journal of International Development 23(4): 461-475.

Fehr, E., et al. (2013). "The development of egalitarianism, altruism, spite and parochialism in childhood and adolescence." European Economic Review 64: 369-383.

Ferraro, P. J. and R. G. Cummings (2007). "Cultural diversity, discrimination, and economic outcomes: an experimental analysis." Economic Inquiry 45(2): 217-232.

Fershtman, C., et al. (2005). "Discrimination and nepotism: The efficiency of the anonymity rule." Journal of Legal Studies 34(2): 371-394.

Fiedler, M., et al. (2011). "Social distance in a virtual world experiment." Games and Economic Behavior 72(2): 400-426.

Filippin, A. and F. Guala (2013). "Costless discrimination and unequal achievements in an experimental tournament." Experimental economics 16(3): 285-305.

Finocchiaro Castro, M. (2008). "Where are you from? Cultural differences in public good experiments." The Journal of Socio-Economics 37(6): 2319-2329.

Fong, C. M. and E. F. Luttmer (2011). "Do fairness and race matter in generosity? Evidence from a nationally representative charity experiment." Journal of Public Economics 95(5): 372-394.

Friesen, J., et al. (2012). "Ethnic identity and discrimination among children." Journal of Economic Psychology 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.

Goette, L., et al. (2006). "The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups." American Economic Review 96(2): 212-216.

Goette, L., et al. (2012). "The Impact of Social Ties on Group Interactions: Evidence from Minimal Groups and Randomly Assigned Real Groups." American Economic Journal-Microeconomics 4(1): 101-115.

Goette, L., et al. (2012). "Competition Between Organizational Groups: Its Impact on Altruistic and Antisocial Motivations." Management Science 58(5): 948-960.

Grossman, P. J. and M. Komai (2008). Leadership and Gender: An Experiment. repository.stcloudstate.edu.

Guala, F., et al. (2013). "Group membership, team preferences, and expectations." Journal of Economic Behavior & Organization 86: 183-190.

Guillen, P. and D. Ji (2011). "Trust, discrimination and acculturation: Experimental evidence on Asian international and Australian domestic university students." The Journal of Socio-Economics 40(5): 594-608.

Gupta, G., et al. (2013). Religion, Minority Status and Trust: Evidence from a Field Experiment, Monash University, Department of Economics.

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?" Journal of Economic Psychology 30(5): 814-827.

Haile, D., et al. (2008). "Cross-racial envy and underinvestment in South African partnerships." Cambridge Journal of Economics 32(5): 703-724.

Hargreaves Heap, S. P. and D. J. Zizzo (2009). "The value of groups." The American Economic Review: 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*." The Economic Journal 121(556): F449-F475.

Hopfensitz, A. and P. Miquel-Florensa (2013). Public good contributions among coffee farmers in costa rica: Cooperativists and private market participants, Toulouse School of Economics (TSE).

Houser, D. and D. Schunk (2009). "Social environments with competitive pressure: Gender effects in the decisions of German schoolchildren." Journal of Economic Psychology 30(4): 634-641.

Ioannou, C. A., et al. (2013). "Group Payoffs As Public Signals." christosaioannou.com.

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." International Journal of Industrial Organization 29(1): 104-115.

Masella, P., et al. (2014). Incentives and group identity. Games and Economic Behavior, 86, 12-25.

McLeish, K. N. and R. J. Oxoby, (2007). Identity, cooperation, and punishment, IZA Discussion Papers, No. 2572

Morita, H. and M. Servátka (2013). Group identity and relation-specific investment: An experimental investigation. European Economic Review, 2013, Vol. 58 (February), pp. 95-109

Netzer, R. J. and M. Sutter (2009). Intercultural trust. An experiment in Austria and Japan, Working Papers in Economics and Statistics.

Ortmann, A. and L. K. Tichy (1999). "Gender differences in the laboratory: evidence from prisoner's dilemma games." Journal of Economic Behavior & Organization 39(3): 327-339.

Pecenka, C. J. and G. Kundhlande (2013). "Theft in South Africa: An Experiment to Examine the Influence of Racial Identity and Inequality." The Journal of Development Studies 49(5): 737-753.

Ploner, M. and I. Soraperra (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. and R. Sosis (2006). "Cooperation and the in-group-out-group bias: A field test on Israeli kibbutz members and city residents." Journal of Economic Behavior & Organization 60(2): 147-163.

Shahriar, Q. (2011). "Identity In A Second ‐ Price Sealed Bid Auction: An Experimental Investigation." The Manchester School 79(1): 159-170.

Slonim, R. and P. Guillen (2010). "Gender selection discrimination: Evidence from a Trust game." Journal of Economic Behavior & Organization 76(2): 385-405.

Solnick, S. J. (2001). "Gender differences in the ultimatum game." Economic Inquiry 39(2): 189-200.

Spiegelman, E. "«C'EST CE QUE JE VOUS DIS»: ESSAIS SUR L'ANALYSE ÉCONOMIQUE DE LA COMMUNICATION INTERPERSONNELLE." (2012).

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." Journal of Economic Psychology 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." International Review of Economics 58(1): 91-103.

Other sources cited:

Aberson, C. L., et al. (2000). "Ingroup bias and self-esteem: A meta-analysis." Personality and Social Psychology Review 4(2): 157-173.

Alesina, A., et al. (2003). "Fractionalization." Journal of Economic 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." Experimental economics 16(2): 170-189.

Arrow, K. (1972). 'Some mathematical models of race discrimination in the labor market', in (A.H. Pascal, ed.), Racial Discrimination in Economic Life, pp. 187–204, Lexington, MA: D.C. Heath.

Balliet, D., et al. (2014). "Ingroup Favoritism in Cooperation: A Meta-Analysis." Psychological Bulletin

Becker, G. S. (2010). The economics of discrimination, University of Chicago press.

Bellemare, C. and S. Kröger (2007). "On representative social capital." European Economic Review 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." Psychological bulletin 127(4): 520.

Blanchard, F. A., et al. (1994). "Condemning and condoning racism: A social context approach to interracial settings." Journal of Applied Psychology 79(6): 993.

Buchan, N. R., et al. (2009). "Globalization and human cooperation." Proceedings of the National Academy of Sciences 106(11): 4138-4142.

Crandall, C. S., et al. (2002). "Social norms and the expression and suppression of prejudice: the struggle for internalization." Journal of personality and social psychology 82(3): 359.

Engel, C. (2007). "How much collusion? A meta-analysis of oligopoly experiments." Journal of Competition Law and Economics 3(4): 491-549.

Engel, C. (2011). "Dictator games: a meta study." Experimental Economics 14(4): 583-610.

Falk, A. and C. Zehnder (2007). Discrimination and in-group favoritism in a citywide trust experiment, IZA Discussion Papers.

Fischer, R. D., 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.

Harbord, R. M., and J. P. Higgins. "Meta-regression in Stata." Meta 8.4 (2008): 493-519.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

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.

Johnson, N. D. and A. A. Mislin (2011). "Trust games: A meta-analysis." Journal of Economic Psychology 32(5): 865-889.

Jones, G. (2008). "Are smarter groups more cooperative? Evidence from prisoner's dilemma experiments, 1959–2003." Journal of Economic Behavior & Organization 68(3): 489-497.

Kiyonari, T., & Yamagishi, T. (2004). In-group cooperation and the social exchange heuristic. In R. Suleiman, D. V. Budescu, I. Fischer, & D. M. Messick (Eds.) Contemporary psychological research on social dilemmas (pp. 269-286). New York, NY: Cambridge University Press.

Oxford Dictionaries, definition of 'political correctness'.
http://www.oxforddictionaries.com/definition/english/political-correctness

Percoco, M. and P. Nijkamp (2009). "Estimating individual rates of discount: a meta-analysis." Applied Economics Letters 16(12): 1235-1239.

Prante, T., et al. (2007). "Evaluating coasean bargaining experiments with meta-analysis." Economics Bulletin 3(68): 1-7.

Robbins, J. M. and J. I. Krueger (2005). "Social projection to ingroups and outgroups: A review and meta-analysis." Personality and Social Psychology Review 9(1): 32-47.

Rosenthal, R. (1979). "The file drawer problem and tolerance for null results." Psychological bulletin 86(3): 638.

Rothstein, H. R., et al. (2006). Publication bias in meta-analysis: Prevention, assessment and adjustments, John Wiley & Sons.

Saucier, D. A., et al. (2005). "Differences in helping whites and blacks: A meta-analysis." Personality and Social Psychology Review 9(1): 2-16.

Tajfel, H., et al. (1971). "Social Categorization and Intergroup Behavior." European Journal of Social Psychology 1(2): 149-177.

Weizsäcker, G. (2010). "Do we follow others when we should? A simple test of rational expectations." The American Economic Review: 2340-2360.

Zizzo, D. J. (2010) "Experimenter demand effects in economic experiments." Experimental Economics 13.1: 75-98.

Appendix 1

Table A1: 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, 2011	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.

Appendix 2: Further Results

Result A1: The strength of discrimination in artificial group experiments does not depend significantly 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 3, Table A3. 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: Country-level variables are not found to significantly 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.

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 3, Table A5, 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[17]) 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

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

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.

Appendix 3: Additional Regression Output

Table A2: Logistic regressions on discrimination and out-group favouritism

Dependent variable	Discrimination	Out-group favouritism
	LOGITa1	LOGITb1
Role Types		
TG Sender	-0.013	0.012
	(0.089)	(0.025)
TG Returner	-0.106	0.029
	(0.079)	(0.030)
Social Dilemma	0.061	0.010
	(0.132)	(0.044)
Allocator	0.422***	-0.030**
	(0.113)	(0.014)
Proposer	-0.052	(dropped)
	(0.105)	
Responder	-0.024	-0.009
	(0.112)	(0.037)
Game Other	0.045	0.027
	(0.107)	(0.041)
Identity		
Ethnicity	-0.254***	0.058
	(0.051)	(0.068)
Religion	-0.188*	-0.025
	(0.103)	(0.055)
Nationality	-0.209***	0.163*
	(0.068)	(0.096)
Gender	-0.242***	0.424***
	(0.044)	(0.103)
Soc/Geo Groupings	0.227**	(dropped)
	(0.103)	
Natural Other	-0.071	(dropped)
	(0.134)	
Controls		
Students	-0.005	-0.037
	(0.071)	(0.057)
Sample Size	2.4e^{-4}	3.0e^{-4}
	(4.7e^{-4})	(2.1e^{-4})
Pseudo R^2	0.171	0.141
N	447	343

Note: *** p<0.01, ** p<0.05, * p<0.1; omitted categories are Dictator (role type) and Artificial (identity); errors are corrected for heteroskedasticity, with 77 clusters in Logit1 and 65 in Logit2; standard errors in parentheses; for dummy variables, dy/dx is for discrete change from 0 to 1.

Table A3: Linear probability regressions on discrimination and meta-regressions on effect size for artificial identity experiments only

Dependent variable	Discrimination	d
	LPM	Metareg
Role Types		
TG Sender	-0.228	-0.148
	(0.203)	(0.146)
TG Returner	-0.167	-0.222
	(0.175)	(0.152)
Social Dilemma	-0.215	0.028
	(0.166)	(0.152)
Allocator	0.236*	0.905***
	(0.138)	(0.170)
Proposer	-0.199	-0.044
	(0.188)	(0.150)
Responder	-0.215	-0.020
	(0.175)	(0.170)
Game Other	-0.073	-0.052
	(0.178)	(0.135)
Controls		
Students	-0.008	0.061
	(0.124)	(0.202)
Sample Size	0.002***	0.001
	(0.001)	(0.001)
Identity Inducement Method		
Labelling	-0.071	0.086
	(0.086)	(0.088)
Other Method	-0.122	0.129
	(0.106)	(0.137)
Team Building	0.124	0.101
	(0.087)	(0.083)
Constant	0.423	0.141
	(0.197)	(0.242)
R^2 (adjusted for Metareg)	0.154	0.262
N	175	150

Note: *** p<0.01, ** p<0.05, * p<0.1; 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.

Table A4: Linear probability regression on discrimination and meta-regression on effect size, with simultaneous and sequential exchange

Dependent variable	Discrimination	d
	LPM	Metareg
Move Sequence		
First Mover	-0.060	0.016
	(0.091)	(0.069)
Second Mover	-0.075	-0.024
	(0.090)	(0.072)
Identity		
Ethnicity	-0.248***	-0.117
	(0.069)	(0.075)
Religion	-0.237*	-0.252**
	(0.138)	(0.123)
Nationality	-0.146	-0.111
	(0.095)	(0.071)
Gender	-0.258***	-0.455***
	(0.085)	(0.126)
Soc/Geo Groupings	0.324***	0.275***
	(0.091)	(0.100)
Natural Other	-0.234	-0.343*
	(0.145)	(0.177)
Controls		
Students	-0.076	-0.046
	(0.058)	(0.084)
Sample Size	4.7e^{-4}	2.3e^{-4}
	(5.7e^{-4})	(5.0e^{-4})
Constant	0.430***	0.295***
	(0.114)	(0.114)
R^2 (adjusted in Metareg)	0.147	0.100
N	327	274

Note: *** p<0.01, ** p<0.05, * p<0.1; omitted categories are simultaneous mover (move sequence) and Artificial (identity); errors in LPM model are corrected for heteroskedasticity, with 63 clusters; standard errors in parentheses.

Table A5: 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
Role Types					
TG Sender	$3.1e^{-4}$	0.172	0.011	0.009	0.009
	(0.109)	(0.090)	(0.028)	(0.027)	(0.027)
TG Returner	-0.147	0.100	0.022	0.017	0.017
	(0.097)	(0.094)	(0.025)	(0.024)	(0.024)
Social Dilemma	-0.069	0.161	0.009	0.009	0.007
	(0.104)	(0.108)	(0.030)	(0.031)	(0.032)
Allocator	0.412***	1.378***	-0.013	-0.032	-0.033
	(0.124)	(0.137)	(0.018)	(0.025)	(0.024)
Proposer	-0.082	0.115	-0.049	-0.038	-0.038*
	(0.124)	(0.115)	(0.031)	(0.033)	(0.033)
Responder	-0.018	0.247*	0.029	0.040	0.040
	(0.123)	(0.135)	(0.054)	(0.055)	(0.055)
Game Other	$2.0e^{-4}$	0.113	0.029	0.043	0.042
	(0.117)	(0.100)	(0.034)	(0.036)	(0.036)
Identity					
Ethnicity	-0.224**	-0.164*	0.049	0.044	0.042
	(0.089)	(0.099)	(0.046)	(0.045)	(0.042)
Religion	-0.152	-0.148	0.008	-0.064	-0.070
	(0.194)	(0.166)	(0.043)	(0.090)	(0.079)
Gender	-0.289***	-0.389***	0.259***	0.251***	0.245***
	(0.098)	(0.103)	(0.046)	(0.047)	(0.044)
Soc/Geo Groupings	0.246**	0.252***	-0.027*	-0.046*	-0.048**
	(0.094)	(0.088)	(0.015)	(0.027)	(0.023)
Natural Other	-0.011	-0.281	-0.029	-0.100	-0.105*
	(0.183)	(0.200)	(0.018)	(0.068)	(0.057)
Controls					
Fractionalisation	0.058	0.318			
	(0.261)	(0.240)			
LogGDPpc	0.016	-0.041			
	(0.086)	(0.083)			
Europe	0.083	0.158			
	(0.126)	(0.141)			
USA	0.036	-0.001	-0.046**	-0.011	
	(0.136)	(0.154)	(0.021)	(0.034)	
Individualism	-0.001	0.002		-0.002	-0.002*
	(0.004)	(0.003)		(0.002)	(0.001)
Constant	0.228	0.218	0.026	0.157	0.168*
	(0.779)	(0.760)	(0.020)	(0.117)	(0.092)
R^2 (adjusted in Metareg1)	0.222	0.357	0.116	0.125	0.125
N	351	308	365	351	351

Note: *** p<0.01, ** p<0.05, * p<0.1; 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.

Appendix 4: Linear Restriction Tests

Table A6: Linear Restriction Tests on models presented in Table 2

Null Hypothesis	P Value on two-tailed test			
	LPMa1	LPMb1	Metareg1	LPMa2
TG Sender = TG Returner	0.269	0.195	0.128	0.525
TG Sender = Social Dilemma	0.585	0.612	0.914	0.818
TG Sender = Allocator	<0.001***	0.002***	<0.001***	0.28
TG Sender = Proposer	0.711	0.331	0.831	0.032**
TG Sender = Responder	0.939	0.248	0.366	0.75
TG Sender = Game Other	0.52	0.443	0.667	0.764
TG Returner = Social Dilemma	0.203	0.169	0.251	0.532
TG Returner = Allocator	<0.001***	<0.001***	<0.001***	0.187
TG Returner = Proposer	0.663	0.909	0.39	0.016**
TG Returner = Responder	0.576	0.064*	0.087*	0.53
TG Returner = Game Other	0.144	0.105	0.446	0.891
Social Dilemma = Allocator	0.010***	0.006***	<0.001***	0.516
Social Dilemma = Proposer	0.397	0.184	0.913	0.060*
Social Dilemma = Responder	0.613	0.481	0.357	0.906
Social Dilemma = Game Other	0.974	0.812	0.772	0.612
Allocator = Proposer	<0.001***	<0.001***	<0.001***	0.114
Allocator = Responder	0.002***	0.069*	<0.001***	0.714
Allocator = Game Other	0.006***	0.012**	<0.001***	0.206
Proposer = Responder	0.859	0.091*	0.338	0.164
Proposer = Game Other	0.264	0.050*	0.873	0.020**
Responder = Game Other	0.499	0.628	0.258	0.588
Ethnicity = Religion	0.694	0.757	0.792	0.423
Ethnicity = Nationality	0.509	0.122	0.716	0.256
Ethnicity = Gender	0.895	0.673	0.004***	<0.001***
Ethnicity = Soc/Geo Groupings	<0.001***	<0.001***	<0.001***	0.072*
Ethnicity = Natural Other	0.223	0.952	0.495	0.139
Religion = Nationality	0.966	0.428	0.615	0.121
Religion = Gender	0.656	0.621	0.053*	<0.001***
Religion = Soc/Geo Groupings	0.001***	0.001***	<0.001***	0.592
Religion = Natural Other	0.447	0.795	0.63	0.843
Nationality = Gender	0.365	0.025**	0.001***	0.082*
Nationality = Soc/Geo Groupings	<0.001***	0.005***	<0.001***	0.017**
Nationality = Natural Other	0.416	0.381	0.412	0.013**
Gender = Soc/Geo Groupings	<0.001***	<0.001***	<0.001***	<0.001***
Gender = Natural Other	0.221	0.912	0.331	<0.001***
Soc/Geo Groupings = Natural Other	0.099*	0.006***	0.003***	0.595

Table A7: Linear Restriction Tests on models presented in Table 4

	P Value on two-tailed test			
Null Hypothesis	**LPMa1**	**Metareg1**	**LPMa2**	**Metareg2**
Ethnicity = Religion	0.168	0.367	0.789	0.9
Ethnicity = Nationality	0.497	0.461	0.439	0.574
Ethnicity = Gender	0.424	0.099*	<0.001***	0.009***
Ethnicity = Soc/Geo Groupings	<0.001***	0.040**	<0.001***	<0.001***
Ethnicity = Natural Other	0.024**	0.934	0.904	0.473
Religion = Nationality	0.099*	0.831	0.936	0.825
Religion = Gender	0.096*	0.621	0.445	0.047**
Religion = Soc/Geo Groupings	<0.001***	0.005***	0.031**	0.003***
Religion = Natural Other	0.011**	0.692	0.874	0.446
Nationality = Gender	0.84	0.335	0.057*	0.003**
Nationality = Soc/Geo Groupings	0.016**	0.020**	<0.001***	<0.001***
Nationality = Natural Other	0.031**	0.794	0.742	0.347
Gender = Soc/Geo Groupings	0.026**	0.002***	<0.001***	<0.001***
Gender = Natural Other	0.037**	0.492	0.418	0.377
Soc/Geo Groupings = Natural Other	0.548	0.322	0.001***	0.002