Culture and cooperation

Simon Gächter1,2,*, Benedikt Herrmann1 and Christian Thöni3

1Centre of Decision Research and Experimental Economics, School of Economics, University of Nottingham, Sir Clive Granger Building, University Park, Nottingham NG7 2RD, UK
2Center for Economic Studies and Ifo Institute for Economic Research, Institute for the Study of Labor, Munich DE-81679, Germany
3University of St Gallen, SEW-HSG, Varnhelsstrasse 14, CH-9000 St Gallen, Switzerland

Does the cultural background influence the success with which genetically unrelated individuals cooperate in social dilemma situations? In this paper, we provide an answer by analysing the data of Herrmann et al. (2008a), who studied cooperation and punishment in 16 subject pools from six different world cultures (as classified by Inglehart & Baker (2000)). We use analysis of variance to disentangle the importance of cultural background relative to individual heterogeneity and group-level differences in cooperation. We find that culture has a substantial influence on the extent of cooperation, in addition to individual heterogeneity and group-level differences identified by previous research. The significance of this result is that cultural background has a substantial influence on cooperation in otherwise identical environments. This is particularly true in the presence of punishment opportunities.

Keywords: human cooperation; punishment; culture; experimental public good games

1. INTRODUCTION

Many important social problems of mankind—from interactions in the workplace to tackling climate change—involve the cooperation of genetically unrelated individuals in situations in which collective welfare is jeopardized by individual self-interest. According to one model of human social behaviour, self-interest is a dominant behavioural force and therefore welfare-enhancing cooperation is doomed to fail, unless well-defined small groups interact indefinitely (which allows for targeted punishment by withdrawing cooperation, see Axelrod (1984), Fudenberg & Maskin (1986), Sigmund (2010)). Numerous behavioural experiments, as for example surveyed in Fehr et al. (2002b), Fehr & Fischbacher (2004) and Gächter & Herrmann (2009), and other empirical studies (e.g. Gintis et al. 2005) have shown that this prediction is way too pessimistic and that much more cooperation exists than is easily compatible with the selfishness assumption. This is particularly true in the presence of punishment opportunities. Many people are willing to exert costly punishment of people whose behaviour they dislike, even when there is no material benefit whatsoever for doing so. However, recent research also suggests that there is substantial individual heterogeneity with regard to prosocial behaviour, in particular in the context of cooperation (e.g. Fischbacher et al. 2001; Kurzban & Houser 2005; Bardsley & Moffatt 2007; Kocher et al. 2008; Muller et al. 2008; Herrmann & Thöni 2009; Fischbacher & Gächter 2010; see Bergmuller et al. 2010, for a discussion of personality and cooperation). Such inter-individual differences have a potential of explaining aggregate behaviour and group-level differences (e.g. Gächter & Thöni 2005; Kurzban & Houser 2005; Gunnthorsdottir et al. 2007; Fischbacher & Gächter 2010) and may play a major role for the stability of cooperation (McNamara & Leimar 2010).

An interesting next approach is to jump from the ‘micro-level’ to the ‘macro-level’ and to ask whether there are also differences in cooperation behaviour across cultural backgrounds. When we speak of the ‘cultural background’ we have in mind those sets of beliefs and values that the majority of people in these societies hold and that get ‘transmitted fairly unchanged from generation to generation’ (Guiso et al. 2006, p. 23). In particular, influential social scientists such as Inglehart (1997) and Inglehart & Baker (2000) argue on the basis of data from the World Values Survey that there are distinct cultural areas in the world, reflected in people’s value systems. The question we ask in this paper is whether there are differences in experimentally observed cooperation behaviour across distinct world cultures.

To answer this question, we will analyse a dataset of highly comparable cross-cultural experiments conducted by Herrmann et al. (2008a) with more than 1100 participants in 16 subject pools from six distinct cultural areas around the world. All participants played finitely repeated public good experiments with and without punishment in stable groups, in a design inspired by Fehr & Gächter (2000). This dataset, which we describe in §2 in more detail, along with our methodology of classifying subject pools according to cultural areas, allows us to disentangle the relative importance of individual heterogeneity, group-level differences and cultural heterogeneity for cooperation. To our knowledge, such an analysis has not been done before.

* Author for correspondence (simon.gaechter@nottingham.ac.uk).
In principle, survey methods could also be applied to uncover cross-cultural differences. However, subjects do not have an incentive to admit their true social preferences when it costs nothing to pass for being cooperative and prosocial. When surveyed, presumably only a few people would admit to being selfish. By contrast, behavioural experiments have the advantage that actual behaviour rather than stated intentions is observed. In experiments participants can, depending on their decisions, earn considerable amounts of money. Thus, the laboratory allows observation of real decision-making under controlled circumstances. Moreover, our goal of disentangling individual heterogeneity, group-level differences and cultural variation demands a laboratory experimental approach.\(^2\)

Why might the cultural background matter at all for cooperation? This is an interesting question because the *Homo economicus* model mentioned above suggests that cultural background does not matter: selfishness is universal. The fact that not all people are selfish has recently inspired theoretical models of social preferences, which take this heterogeneity into account (e.g. Fehr & Schmidt (2006) for a survey). Yet, these models are also mute with respect to the influence of the cultural background. In general, economists, with some exceptions (e.g. Roth et al. (1991) in a seminal study) have not been interested in cultural differences. This is now changing (for a succinct survey see Fernández 2008). The reasons are theoretical developments (e.g. Greif 1994; Bowles 1998; Bednar & Page 2007; Guiso et al. 2008; Tabellini 2008b), and better data, both experimental (e.g. Henrich et al. 2001; Oosterbeek et al. 2004) and non-experimental (Guiso et al. 2006; Fernández 2007; Tabellini 2008a). By contrast, psychologists have established many profound differences in human behaviour and thinking across cultures (e.g. the reviews by Markus & Kitayama (1991), Nisbett & Cohen (1996), Cohen (2001), Nisbett (2003), Henrich et al. (in press), Heine & Buchtel (2009) and Heine & Ruby (2010)). For example, in a recent paper Henrich et al. (in press) show that Western subjects, who are most frequently used in behavioural experiments, are actually often the outlier in the range of observed behaviours. Thus, it is an obvious question whether there are also differences in cooperation behaviour across different world cultures. Moreover, evolutionary psychological approaches predict the possibility of cultural differences because people have an evolved psychology that allows them to attain their behaviour to the norms, expectations and (sanctioning) behaviours of others around them (e.g. Boyd & Richerson 2005; Henrich 2004; Henrich & Henrich 2007; Herrmann et al. 2007; Nettle 2009; Tomasello et al. 2005; Tomasello 2009; Rendell et al. 2010; Gintis in press).

From what we know from numerous experiments, we can speculate about potential behavioural channels of cultural influences. First, in the context of cooperation many experiments have shown that people are conditional cooperators who cooperate more the more they believe others will cooperate (e.g. Croson 2007; Gächter 2007; Fischbacher & Gächter 2010). Any factor that influences beliefs might also influence cooperation. This is also true of framing effects (e.g. Dufwenberg et al. 2006) or, more generally, contextual cues, of which the cultural background is an important example. For example, subjects in a public good experiment in Kenya termed the neutrally framed experiment as ‘harambee’, their word for community work (Henrich et al. 2005). The way naturally occurring cooperation problems are normally solved in society might influence people’s beliefs about how others will behave. Second, from experiments in which punishment was possible, we know that substantial differences in punishment across subject pools in different cultures can exist and even be anticipated prior to any experience in the particular situation (e.g. Gächter et al. (2005) and Gächter & Herrmann (2009) who ran experiments in Russia and Switzerland). Consistent with this observation, Herrmann et al. (2008a), in experiments which we shall analyse in detail below, showed a large diversity of punishment patterns across different subject pools around the world, resulting in vastly different cooperation levels.

We are of course not the first to investigate cultural influences on cooperation behaviour or prosociality in general (e.g. Oosterbeek et al. 2004). Particularly noteworthy are the seminal large-scale studies conducted in small-scale societies around the world (Henrich et al. 2010).\(^3\) While Henrich and his co-workers (Henrich et al. 2001, 2005, 2006) mostly used simple bargaining games and conducted their experiments with members of small-scale societies, the experiments we shall analyse were all conducted in large-scale developed societies. The small-scale societies differ among each other in the extent to which cooperation is important for economic production (e.g. cooperative whale-hunting versus individual hunting and gathering); how strong market integration is (how many calories are bought on the market?); the size of communities; and adherence to a world religion (Henrich et al. 2010). Differences on these dimensions explain a large part of the variation that is observed in experimental bargaining games in these small-scale societies (Henrich et al. 2005, 2010). Modern developed societies hardly differ on the dimensions of market integration and reliance on cooperation, for all modern societies know division of labour and trade between non-kin (Richerson & Boyd 1999). Thus, in comparison with the small-scale societies, the cultural influence we identify does not come from fundamental differences in socio-economic structures but from historical, religious, political and value differences, which Inglehart & Baker’s (2000) classification of cultural areas around the world, or Hofstede’s (2001) ‘cultural dimensions’ try to capture.

Another distinguishing feature of our approach from previous cross-cultural economics experiments is that many of them test specific (proximate) hypotheses that are derived from the compared cultures (Yamagishi 1988; Yamagishi & Yamagishi 1994; Kachelmeier & Shehata 1997; Yamagishi et al. 1998; Hayashi et al. 1999; Buchan et al. 2002, 2009; Holm & Danielson 2005; Chuah et al. 2007, 2009; Bohnet et al. 2008, 2010; Wu et al. 2009; Bornhorst et al. in press).
Our approach is different since our goal is to understand a more fundamental issue—do we find evidence that comparable subjects from modern developed societies that are characterized by large-scale cooperation but differ strongly with regard to historical and cultural values behave differently in games of cooperation? This question is motivated by evolutionary theories of cooperation (Sober & Wilson 1998; Henrich 2004; Nowak 2006; Henrich & Henrich 2007) rather than proximate mechanisms of cultural differences.

The typical methodology of cross-cultural experiments is to observe a comparable subject pool in different societies. The idea is to run experiments in a way that minimizes variations owing to subject pool composition or experimental procedures. In this way any differences that might be observed between cross-societal subject pools are probably due to differences in the cultural background of the compared societies. Our methodology, which we describe in more detail in the next section, builds on this idea but refines it in two ways. First, the data of Herrmann et al. (2008a) were collected in six distinct cultural areas according to Inglehart & Baker (2000) and Hofstede (2001). Thus, rather than comparing two cultures, we compare six cultures. Second, we do not identify culture by nationality, because different nations can share largely similar cultural backgrounds. The cultural classification of Inglehart & Baker (2000) gives us at least two different societies in each of the six cultures; in three cultural areas we have data from subject pools from three different societies and in three cultural areas from two different societies. In one culture, ‘Protestant Europe’, we have data from four subject pools from three countries (in Switzerland we have data from two subject pools, St Gallen and Zurich). This structure of our data allows us to compare within-cultural variation with between-cultural variation, which is impossible if there is only one subject pool per society or cultural area.4 Our main findings are that cooperation within cultures is largely similar while there exist highly significant differences between cultures. This is true in public good experiments with and without punishment and also holds for punishment behaviour. This dual observation of within-culture similarity and cross-cultural heterogeneity is the main support for the claim that there are cultural influences on cooperation.

2. THE DATA AND OUR APPROACH
In the following, we first describe the most important details of the design of Herrmann et al. (2008a), followed by the details of our classification of cultural areas. Our third step is a description of our main statistical approach for discerning the importance of the cultural background for cooperation and punishment.

We start with the details of the experimental design, which was motivated by the observations from Ostrom et al. (1992), Fehr & Gächter (2000) and Fehr & Gächter (2002) who showed that the punishment mechanism has dramatic impacts on contributions in the public goods game. All subjects took part in two experiments, each lasting for 10 periods. The first experiment always was a public good experiment with no punishment opportunities (we call this the ‘N-condition’). The second experiment was a public good experiment with a punishment opportunity (the ‘P-condition’). Both experiments were played in stable groups of four subjects for 10 periods. In both experiments, subjects received an endowment of 20 ECU (experimental currency unit) in each period. All subjects decided simultaneously how many ECU they wanted to contribute to a public good. All contributions in a group were summed up and multiplied by 1.6. The resulting amount of ECU was divided equally among all subjects in the group. A subject’s payoff consisted of the ECUs he or she did not contribute plus his share of the public good. In the N-condition, the stage game ended here and subjects moved on to the next period. Note that in this game it is individually rational (assuming selfish preferences) to contribute nothing to the public good: for every unit contributed a subject earns only 0.4 units in return. However, joint income is maximized if all subjects contribute their entire endowment to the public good. This is due to the fact that the social return of contributing is 1.6 per unit contributed.

In the P-condition, there was an additional stage where subjects could reduce each other’s incomes at their own cost. All subjects learned the contributions of all other group members. Subjects could then assign punishment points to each other group member. Each punishment point reduced the income of the punished group member by three ECUs. However, punishment was also costly to the punisher. Each punishment point cost the punisher one ECU. For further details, the procedures and the instructions, we refer the reader to Herrmann et al. (2008b).

Herrmann et al. (2008a) ran these experiments in 16 different locations with a total of 1120 participants. The locations are all over the developed world and span a large set of cross-societal differences (see Herrmann et al. (2008b) for the details). As explained above, subjects interacted in stable groups of four members throughout the entire experiment. Therefore, groups constitute the independent units of observations on which all our non-parametric tests will be based. In total, we have data from 280 groups.

Herrmann et al. (2008a) designed and ran their experiments in a way that minimizes differences in behaviour that come from subject pool composition or experimental procedures. To ensure this, participants were all undergraduates and thereby very similar with regard to age, education and their socio-economic situation in their respective society. Gender composition was also similar in most subject pools. Thus, any variation we observe between subject pools or cultural regions are unlikely owing to differences in subject pool composition. Similarly, to minimize behavioural variability as introduced by experimental procedures, Herrmann et al. (2008a) followed standard practices of cross-cultural experiments as introduced to experimental economics by Roth et al. (1991). A detailed discussion of these issues can be found in Herrmann et al. (2008b).

An important conceptual step for our purposes is to classify locations into cultural regions according to
cultural proximity. To avoid being arbitrary, we rely on seminal research by Inglehart (1997) and Inglehart & Baker (2000), who used data from the World Values Survey to identify clusters in world cultures. According to Inglehart & Baker (2000), societies can be characterized by two dimensions: ‘traditional versus secular-rational values’ and ‘survival versus self-expression values’. The first refers to people’s attitudes on topics like abortion, national pride, obedience and respect for authorities; the latter refers to attitudes on the importance of economic and physical security over self-expression and quality-of-life; homosexuality, happiness and trust. Table 1 shows the countries where our data stems from and their cultural classification. Where available, we take the classification from the Global Cultural Map (Inglehart & Baker 2000, p. 29 fig. 1). This allows us to classify all countries in the cultural areas ‘English speaking’, ‘Protestant Europe’, ‘Orthodox/ex-Communist’ and ‘Confucian’. Among the four remaining countries, only Turkey appears in Inglehart & Baker (2000). An alternative source of information about cultural differences are the four cultural dimensions (power distance, individualism, masculinity and uncertainty avoidance) defined by Hofstede (2001). Using these four dimensions strongly suggests pairing Greece and Turkey. If we calculate the Euclidian distance, then Turkey is the third closest country to Greece in a sample of 71 countries (and the closest one in our sample of countries). Finally, we group the two Arabic subject pools into the category ‘Arabic-speaking’.

Before we continue, a caveat is in order. Classifications are always to some extent open to criticism, and Inglehart & Baker (2000) are aware of this (see their discussion on pp. 32–40). We believe, however, that this classification makes a lot of sense, in particular because the identified cultural clusters all share some common history and four of the clusters also share a common language. Moreover, the identified clusters are also similar with regard to other measures of cultural similarity, such as Hofstede’s four cultural dimensions (Hofstede 2001), or norms of civic cooperation, the strength of the rule of law or democracy (see Herrmann et al. 2008b, in particular table S1). There is no detailed information on the Arabic countries, but Hofstede groups them under ‘Arab world’ (Hofstede’s sample does not include Oman, but its neighbouring states Saudi Arabia and UAE).

Our main interest is in whether there are cultural differences in contribution decisions and how important they are, if they exist. To analyse these questions our empirical strategy will be twofold. We first describe the data using graphical tools and non-parametric tests to analyse whether there are cultural differences, that is, systematic patterns of different contributions to the public good according to the cultural areas defined above. Cultural differences exist whether the variation between cultures is larger than the variation within cultures. Therefore, we will provide tests of behaviour within a culture as well as tests between cultures. If behaviour is very homogeneous within the culture but different across cultures, we should not find statistically significant differences within the culture but significant differences between cultures. Notice, however, that homogeneity within the culture and differences across cultures are only sufficient for the existence of cultural differences. Significant between-cultural differences can still exist even if there are significant within-cultural differences, provided the within-cultural differences are ‘small enough’ relative to the between-cultural differences.

The existence of cultural differences does not yet tell us how ‘big’ they are, also relative to the importance of individual variation and variation that is due to differences between groups. For that purpose we use a nested analysis of variance (ANOVA) model to attribute the amount of variance in the contributions explained by cultural variations, group differences and individual heterogeneity. Our basic linear model underlying the ANOVA uses the exogenous variables Period, Culture, Group and Individual (Period is the period number, Culture is a categorical variable to identify the six cultural clusters, and Individual (Group) is a dummy variable for each individual (group)). Individual is nested in Group and Group is nested in Culture. We use the ANOVA to disentangle the coefficient of determination to separate the explanatory power of our exogenous variables in the N- and the P-condition. Our approach not only allows us to measure the explanatory power of cultural variation, but also allows us to compare the importance of cultural variation relative with individual and group influences.

### Table 1. Cultural classification and number of observations of the cities where our data stems from. Classification taken and adapted from Inglehart & Baker (2000) and Hofstede (2001) (for Southern Europe and Arabic speaking).

<table>
<thead>
<tr>
<th>city</th>
<th>country</th>
<th>culture</th>
<th>no. of subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>USA</td>
<td>English speaking</td>
<td>56</td>
</tr>
<tr>
<td>Nottingham</td>
<td>UK</td>
<td></td>
<td>56</td>
</tr>
<tr>
<td>Melbourne</td>
<td>Australia</td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>Copenhagen</td>
<td>Denmark</td>
<td>Protestant Europe</td>
<td>68</td>
</tr>
<tr>
<td>Bonn</td>
<td>Germany</td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>Zurich</td>
<td>Switzerland</td>
<td></td>
<td>92</td>
</tr>
<tr>
<td>St Gallen</td>
<td>Switzerland</td>
<td></td>
<td>96</td>
</tr>
<tr>
<td>Minsk</td>
<td>Belarus</td>
<td>Orthodox/</td>
<td>68</td>
</tr>
<tr>
<td>Dnipropetrovsk’k</td>
<td>Ukraine</td>
<td>ex-Communist</td>
<td>44</td>
</tr>
<tr>
<td>Samara</td>
<td>Russia</td>
<td></td>
<td>152</td>
</tr>
<tr>
<td>Athens</td>
<td>Greece</td>
<td>southern Europe</td>
<td>44</td>
</tr>
<tr>
<td>Istanbul</td>
<td>Turkey</td>
<td></td>
<td>64</td>
</tr>
<tr>
<td>Riyadh</td>
<td>Saudi Arabia</td>
<td>Arabic speaking</td>
<td>48</td>
</tr>
<tr>
<td>Muscat</td>
<td>Oman</td>
<td></td>
<td>52</td>
</tr>
<tr>
<td>Seoul</td>
<td>South Korea</td>
<td>Confucian</td>
<td>84</td>
</tr>
<tr>
<td>Chengdu</td>
<td>China</td>
<td></td>
<td>96</td>
</tr>
</tbody>
</table>

3. RESULTS

The main results of the first part of our analysis, which concerns the existence of culture effects, are contained in figures 1–4. Recall that we argued that cultural differences in contributions exist if contributions are more similar within a culture than between cultures. In our analysis, we separate the data according to the
games. To test whether there are also cultural differences with regard to the extent of the decline of cooperation, we calculated for each independent group a Spearman rank order correlation of group average contribution and period. We use this correlation coefficient as a test statistic in a Kruskal–Wallis test with the cultural regions as the test groups. We find highly significant differences ($\chi^2(5) = 42.1, p = 0.0001$).

We now turn to the analysis of the P-condition (illustrated in the right part of each panel). Within a culture the temporal patterns are surprisingly similar. In some of the cultures there is also an indication of significant within-culture variation: cooperation levels are significantly different in two and weakly significantly different in one culture. Across cultures cooperation levels are highly significantly different (Kruskal–Wallis test with group averages as independent observations and culture as the grouping variable; $\chi^2(5) = 96.5, p = 0.0001$).

Figure 1 (and figures 3 and 4 below) also suggest that there are cultural differences with regard to the change of contributions between the N-condition and the P-condition: in four cultures contributions are significantly higher in the P-condition than in the N-condition (with $p < 0.002$) whereas in two cultures this change is not significant (with $p > 0.459$, Wilcoxon signed-ranks tests with group averages as independent observations (see the $p$-values for ‘change’ indicated in figure 1).

---

**Figure 1.** Average contributions in the 16 subject pools during the 10 periods of the N-condition and the P-condition; ‘c’ denotes the average contribution across all periods and subject pools of a given treatment and culture; ‘p’ denotes the $p$-value of a Kruskal–Wallis test for the equality of contributions of subject pools in a given treatment and culture. ‘Change’ denotes the $p$-value of a Wilcoxon signed-rank test for the change of contribution between the N-condition and the P-condition. All tests are based on group average contributions over all periods of a respective treatment.

---

This difference concerns the average level of cooperation. However, all subject pools experience a decline of contributions in the N-condition over time (except subjects in Athens and the two Arabic subject pools, where contributions appear more stable). The explanation of the decline of cooperation is beyond the scope of this paper. We refer the reader to Neugebauer et al. (2009) and Fischbacher & Gächter (2010) for analyses of the almost ubiquitous decline of cooperation in finitely repeated public good games. To test whether there are also cultural differences with regard to the extent of the decline of cooperation, we calculated for each independent group a Spearman rank order correlation of group average contribution and period. We use this correlation coefficient as a test statistic in a Kruskal–Wallis test with the cultural regions as the test groups. We find highly significant differences ($\chi^2(5) = 42.1, p = 0.0001$).

We now turn to the analysis of the P-condition (illustrated in the right part of each panel). Within a culture the temporal patterns are surprisingly similar. In some of the cultures there is also an indication of significant within-culture variation: cooperation levels are significantly different in two and weakly significantly different in one culture. Across cultures cooperation levels are highly significantly different (Kruskal–Wallis test with group averages as independent observations and culture as the grouping variable; $\chi^2(5) = 96.5, p = 0.0001$).

Figure 1 (and figures 3 and 4 below) also suggest that there are cultural differences with regard to the change of contributions between the N-condition and the P-condition: in four cultures contributions are significantly higher in the P-condition than in the N-condition (with $p < 0.002$) whereas in two cultures this change is not significant (with $p > 0.459$, Wilcoxon signed-ranks tests with group averages as independent observations (see the $p$-values for ‘change’ indicated in figure 1).
We conclude from this analysis that there are cultural differences in contributions, in particular in the P-condition. The major part of these cultural differences in the P-condition is most probably owing to differences in punishment. Antisocial behaviour increasingly attracts attention in the study of cooperation (Jensen 2010) but the role of culture remains little explored. Herrmann et al. (2008), table 1, show that contributions are strongly linked to patterns of punishment. In particular they show that contributions in the P-condition depend (i) positively on the initial contribution, (ii) positively on the extent of punishment of free-riding behaviour, and (iii) negatively on antisocial punishment, that is punishment of people who contributed the same or more than the punishing individual. Herrmann et al. (2008a) also show that antisocial punishment is strongly linked to norms of civic cooperation in a given society as measured by representative questionnaires in the World Values Survey and the strength of the rule of law in a country (see Herrmann et al. (2008b) for further details and references). Both measures differ strongly between the societies of the subject pools of Herrmann et al. (2008a). Thus, (antisocial) punishment seems to be linked to societal background. This observation begs the question of cultural differences in punishment behaviour. Herrmann et al. (2008a) have already shown that there are only weakly significant differences in punishment of free-riding behaviour and highly significant differences in antisocial punishment across subject pools. Are there cultural differences in punishment if we apply our concept of cultural differences?

Figure 2 depicts the extent of average punishment of free-riding behaviour as well as of antisocial punishment per subject pool and grouped for the six cultural areas. Interestingly, with one exception, there are no significant differences in both free rider punishment and antisocial punishment within cultures (based on Kruskal–Wallis tests). Moreover, we find significant differences in punishment across cultures for free rider punishment ($\chi^2(5) = 11.2, \ p = 0.048$) and much stronger cultural differences in antisocial punishment ($\chi^2(5) = 82.5, \ p = 0.0001$).

In addition to the culture-specific changes in contributions between conditions, figures 3 and 4 illustrate two further features of the data, which we shall analyse in more detail in the next step. Figure 3 focuses on the distribution of individual average contributions and shows that in the cultures in which punishment leads to a significant behavioural change, the variance of individual contributions is reduced as well. Not very surprisingly, punishment, when it ‘works’, makes people’s contributions more similar (and increases the level of contributions), whereas no such homogenizing effect is visible when punishment is ineffective. In two cultures, the variance of individual contributions even increases in the presence of punishment.

Figure 4 illustrates how group average contributions are distributed between conditions and cultures. This is interesting because cooperation in the Herrmann et al. (2008a) experiments happened in groups...
Figure 3. Histograms of individual average contributions in the N- and P-condition for each culture. The numbers in each panel indicate the standard deviation of the contributions in a culture in the two conditions. To measure the standard deviation of contributions independently of the time trend, we calculate a standard deviation for each of the 10 periods and report the average standard deviation across the 10 periods.

Figure 4. Histograms of average group contributions per period in the N- and P-condition for each culture. The numbers indicate standard deviations of the contributions within a group and between groups. For the within-group standard deviations, we calculate the standard deviations of the four contributions in a group in each period and average over all periods and groups within a culture. For the between-group measure, we calculate the standard deviation of all group averages within a culture and a period. The numbers show the average over the 10 periods for the N- and the P-condition.
with fixed memberships over time and groups might have been ‘locked’ into a particular path-dependent contribution pattern, for example owing to a frequent tendency of conditional cooperation (e.g. Gächter & Thöni 2005; Kurzban & Houser 2005; Gunnthorsdottir et al. 2007; Fischbacher & Gächter 2010). Such path-dependency might lead to substantially different group average contributions, and therefore to large between-group variance. Moreover, the presence of punishment might affect both the between-group variance (by making groups more homogeneous) and the within-group variance. We find that the introduction of punishment reduces the within-group variance in all six cultures. The effect on the between-group variance is more diverse: in four of the six cultures the between-group variance increases and in two cultures it decreases.

We conclude from this descriptive analysis that cultural differences in contribution decisions exist without doubt. In our next step, we are interested in the relative fraction of the variance that is due to individuals, groups and in particular culture in contributions in both the N- and the P-conditions. For this purpose, we use the nested ANOVA model described in §2 to decompose the explanatory power of our measure for culture, group composition and individual differences.

Figure 5 shows the $R^2$-associated to our explanatory variables for the N- and the P-conditions. It measures the sum of squares associated with the explanatory variable divided by the total sum of squares in the contribution decisions. Bar heights depict the fraction of the variance that is explained by the corresponding variable. The lowest part of a bar depicts the fraction of the variance explained by Culture. In the N-condition, the cultural variation in our subject pool explains only a small amount of the variance (3.9%). Group-level differences (that is, between-group variance) account for additional 29.3 per cent of the variation in contributions, and a further 16.0 per cent can be explained by individual-fixed effects. Time effects account for 7.4 per cent of the variation. Finally, 43.4 per cent of the variation remains unexplained by our model.

Comparing the results of the N-condition to the results of the P-condition reveals striking differences. First, a much smaller portion of the variance in contributions remains unexplained. Individual and period effects lose much of their explanatory power while Group and Culture gain in importance. In particular, the percentage of the variance explained by our cultural classification is more than five times larger in the P-condition than in the N-condition.

Are these fractions of explained variance large? This is an important question, because even in the absence of any systematic cultural, group or individual effects the ANOVA model would provide some non-zero $R^2$. We ran 100 ANOVAs with simulated contributions (all contributions in 0,1,…,20 were drawn with equal probability). The explanatory power of Culture in the absence of systematic cultural variation is very close to zero (mean: 0.043%, s.d. 0.025). Consequently, the influence of Culture is far beyond the effect that would show up in the absence of cultural variation. The same is true for Group and Individual effects, as well as Period effects.  

4. SUMMARY AND CONCLUDING REMARKS

In this paper, we have analysed an experimental dataset by Herrmann et al. (2008a) who ran comparable public good experiments with and without punishment in 16 subject pools from six distinct cultural areas around the world. This dataset allows us to show that cultural differences in cooperation exist in the sense that within-cultural variation is smaller than the between-cultural variation. Moreover, we found that for the extent of cooperation we observe, cultural variation is a particularly important source of variation in the presence of punishment opportunities. This is due to large cultural differences in punishment. In the absence of punishment, individual (‘micro-level’) variation is much more important than cultural (‘macro-level’) variation, whereas the opposite is true in the presence of punishment. Group-level differences (the ‘meso-level’) are very important both in the presence and the absence of punishment.

We know from numerous experiments that individual heterogeneity is an important source of variation that can translate into important aggregate-level differences in outcomes (Camerer & Fehr 2006; Gächter & Thöni in press). Our dataset confirms this insight by showing that individual variation and group-level variation are both important sources of the overall variation. The importance of our finding of culture effects in addition to individual-level and group-level differences is that, holding everything else constant, differences in cultural background can lead to differences in behaviour in otherwise identical environments. Thus, accounting for individual and implied group-level differences is not enough to understand the whole breadth of variation in cooperation. Culture needs to be accounted for.

We conclude with two caveats and future research questions. First, in this analysis, we have only demonstrated the existence and quantitative importance of cultural differences. Our approach cannot explain where the differences come from. Herrmann et al. (2008a) found large differences in cooperation only in the presence of punishment and owing to large
differences in punishment across subject pools. Ascertaining why these cultural differences in punishment occur is an interesting task for future research. Second, we have drawn our conclusions from comparing subjects who are very similar with regard to their socio-economic status and other socio-demographic characteristics. However, in every society there exist various social groups who might also show a large variation in cooperative behaviour (e.g. Ockenfels & Weimann 1999; Fehr et al. 2002a; Bellemare & Kröger 2007; Hong & Bohnet 2007; Hoff et al. 2009; Kocher et al. 2009; Gächter & Herrmann in press; Henrich et al. in press). It is an important task for future research to understand this sort of variation relative to the sources of variation we have identified in this paper.

We gratefully acknowledge financial support from the University of Nottingham, the Latsis Foundation (Geneva), and the EU-TMR Research Network ENDEAR (FMX-CT98-0238). We received helpful comments from the editors and referees, as well as various workshop audiences, in particular the Arts and Humanities Research Council workshops Culture and the Mind in Sheffield, and from Peter Egger and Conny Wunsch. This paper is part of the MacArthur Foundation Network on Economic Environments and the Evolution of Individual Preferences and Social Norms.

ENDNOTES

1 Gächter et al. (2004) and Thöni et al. (2009) show that, on the individual level and within a given culture, there is a connection between questionnaire items as used in the World Values Survey and cooperation in public goods games. Such a relationship has also been established in trust games, which also contain an element of cooperation (Ermisch et al. 2009).


3 The most important experimental tool in these studies is the ultimatum game (Güth et al. 1982). For a comprehensive analysis and cross-cultural comparison of ultimatum bargaining games see Oosterbeek et al. (2004).

4 Gächter & Herrmann (2009) applied this methodology to one-shot experiments conducted with students (n = 606) in two Swiss subject pools and two Russian subject pools. According to several measures, Russia and Switzerland are culturally very distinct societies. The results show within-cultural similarity but strong between-cultural differences.

5 In general, the ANOVA does not allow for an unambiguous disaggregation of the coefficient of determination. In our analysis this is possible because all exogenous variables are orthogonal and our sample is balanced.

6 The simulation (n = 100) for Group yields mean = 2.42%, s.d. = 0.197; for Individual mean = 7.51%, s.d. = 0.330. Period: mean = 0.082%, s.d. = 0.039.

REFERENCES


Phil. Trans. R. Soc. B (2010)