Institute for Empirical Research in Economics
University of Zurich

Working Paper Series
ISSN 1424-0459

Forthcoming in:
Foundations of Human Sociality –
Experimental and Ethnographic Evidence from 15 Small-Scale Societies,
edited by Henrich, Boyd, Bowles, Camerer, Fehr, Gintis and McElreath.

Working Paper No. 97

Measuring Social Norms and Preferences using Experimental Games: A Guide for Social Scientists

Colin F. Camerer and Ernst Fehr

January 2002

Measuring social norms and preferences using experimental games: A guide for social scientists

Colin F. Camerer

Division of Humanities and Social Sciences,
California Institute of Technology,
Pasadena, CA 91125
camerer@hss.caltech.edu.

Ernst Fehr

Institute for Empirical Research in Economics
University of Zürich
Blümlisalpstrasse 10, CH – 8006 Zürich
efehr@iew.unizh.ch

May 2001

Forthcoming in: Foundations of Human Sociality – Experimental and Ethnographic Evidence from 15 Small-Scale Societies, edited by Henrich, Boyd, Bowles, Camerer, Fehr, Gintis and McElreath.

This paper was prepared for the MacArthur Foundation Anthropology project meeting. This research was supported by NSF SBR9730364. Thanks to Sam Bowles and Joe Henrich for comments, and Natalie Smith for sharing figures from her paper. Ernst Fehr acknowledges support from the Swiss National Science Foundation (project number 1214-05100.97), the Network on the Evolution of Preferences and Social Norms of the MacArthur Foundation, and the EU-TMR Research Network ENDEAR (FMRX-CTP98-0238).

1. Introduction

The purpose of this chapter is to describe a menu of experimental games that are useful for measuring aspects of social norms and social preferences. Economists use the term "preferences" to refer to the choices people make, and particularly to tradeoffs between different collections ("bundles") of things they value—food, money, time, prestige, and so forth. "Social preferences" refer to how people rank different allocations of material payoffs to themselves and others. Self-interested individuals care only about their own material payoffs. The past two decades of experimental research have shown, however, that a substantial fraction of people in developed countries (typically college students) also care about the payoffs of others. In some situations, many people are willing to spend resources to reduce the payoff of others. In other situations, the same people spend resources to increase the payoff of others.

As we will see, the willingness to reduce or increase the payoff of relevant reference actors exists even though people reap neither present nor future material rewards from reducing or increasing payoffs of others. This indicates that, in addition to self-interested behavior, people sometimes behave as if they have altruistic preferences, and preferences for equality and reciprocity.[1] Reciprocity, as we define it here, is different from the notion of reciprocal altruism in evolutionary biology. Reciprocity means that people are willing to reward friendly actions and to punish hostile actions *although the reward or punishment causes a net reduction in the material payoff of those who reward or punish.* Similarly, people who dislike inequality are willing to take costly actions to reduce inequality although this may result in a net reduction of their material payoff. Reciprocal altruism typically assumes that reciprocation yields a net increase in the material payoff (for example, because one player's action earns them a reputation which benefits them in the future). Altruism, as we define it here, means that an actor takes costly actions to increase the payoff of another actor, *irrespective of the other actor's previous actions*. Altruism thus represents unconditional kindness while reciprocity means non-selfish behavior that is conditioned on the previous actions of the other actor.

[1] We defer the question of whether these *preferences* are a stable trait of people, or tend to depend on situations. While many social scientists tend to instinctively guess that these preferences are traits of people, much evidence suggests that cross-situational *behavior* is not very consistent at the individual level. Note, however, that behavioral variations across situations do not imply that preferences vary across situations because individuals with fixed preferences may well behave differently in different situations (see section 3 below).

Reciprocity, inequality aversion and altruism can have large effects on the regularities of social life and, in particular, on the enforcement of social norms. This is why the examination of the nature of social preferences is so important for anthropology and for social sciences in general. There is, for example, an ongoing debate in anthropology about the reasons for food-sharing in small-scale societies. The nature of social preferences will probably have a large effect on the social mechanism that sustains food-sharing. For example, if many people in a society exhibit inequality aversion or reciprocity, they will be willing to punish those who do not share food, so no formal mechanism is needed to govern food-sharing. Without such preferences, formal mechanisms are needed to sustain food-sharing (or sharing does not occur at all). As we will see there are simple games that allow researchers to find out whether there are norms of food-sharing, and punishment of those who do not share.

In the following we first sketch game theory in broad terms. Then we describe some basic features of experimental design in economics. Then we introduce a menu of seven games that have proved useful in examining social preferences. We define the games formally, show what aspects of social life they express, and describe behavioral regularities from experimental studies. The behavioral regularities are then interpreted in terms of preferences for reciprocity, inequity aversion or altruism. The final sections describe some other games anthropologists might find useful, and draw conclusions.

2. Games and game theory

Game theory is a mathematical language for describing strategic interactions and their likely outcomes. A game is a set of strategies for each of several players, with precise rules for the order in which players choose strategies, the information they have when they choose, and how they rate the desirability (``utility'') of resulting outcomes. Game theory is designed to be flexible enough to be used at many levels of detail in a broad range of sciences. Players may be genes, people, groups, firms or nation-states. Strategies may be genetically-coded instincts, heuristics for bidding on the e-Bay website, corporate routines for developing and introducing new products, a legal strategy in complex mass tort cases, or wartime battle plans. Outcomes can be anything players value-- prestige, food, control of Congress, sexual opportunity, returning a tennis serve, corporate profits, the gap between what you would maximally pay for something and what you actually pay (“consumer surplus”), a sense of justice, or captured territory.

Game theory consists of two different enterprises: (1) Using games as a language or taxonomy to parse the social world; and (2) deriving precise predictions about how players will play in a game by assuming that players maximize expected "utility" (personal valuation) of consequences, plan ahead, and form beliefs about other players' likely actions. The second enterprise dominates game theory textbooks and journals. Analytical theory of this sort is extremely mathematical, and inaccessible to many social scientists outside of economics and theoretical biology. Fortunately, games can be used as a taxonomy with minimal mathematics because understanding prototypical games— like those discussed in this chapter— requires nothing beyond simple logic.

The most central concept in game theory is Nash equilibrium. A set of strategies (one for each player) form an equilibrium if each player is choosing the strategy which is a best response (i.e., gives the highest expected utility) to the other players' strategies. Attention is focussed on equilibrium because players who are constantly switching to better strategies, given what others have done, will generally end up at an equilibrium. Increasingly, game theorists are interested in the dynamics of equilibration as well, in the form of evolution of populations of player strategies (Weibull, 1995); or learning by individuals from experience (e.g., Fudenberg and Levine, 1998; Camerer and Ho, 1999).

Conventions in economic experimentation

At this point, it is useful to describe how experimental games are typically run (see Camerer, in press; Friedman and Sunder, 1993; Davis and Holt, 1995 for more methodological details). Experimental economists are usually interested initially in interactions among anonymous agents who play once, for real money, without communicating. This stark situation is not used because it is lifelike (it's not). It is used as a benchmark from which the effects of playing repeatedly, communicating, knowing who the other player is, and so forth, can be measured by comparison.

In most experiments described below, subjects are college undergraduates recruited from classes or public sign-up sheets (or increasingly, email lists or websites) with a vague description of the experiment (e.g., "an experiment on interactive decision-making") and a range of possible money earnings. The subjects assemble and are generally assigned to private cubicles or as groups to rooms. Care is taken to ensure that any particular subject will not know precisely whom they are playing. If subjects know who they are playing, their economic incentives may be distorted in a way the experimenter does not understand (e.g., they may help friends earn more) and there is an

opportunity for post-game interaction which effectively changes the game from a one-shot interaction to a repeated interaction.

The games are usually described in plain, abstract language, using letters or numbers to represent strategies rather than concrete descriptions like "helping to clean up the park" or "trusting somebody in a faraway place". As with other design features, abstract language is used not because it is lifelike, but as a benchmark against which the effects of concrete descriptions can be measured. It is well-known that there are framing effects, or violations of the principle of description invariance—how the experiment is described may matter. For example, in public goods games players who are asked to take from a common pool for their private gain typically behave differently than subjects who are asked to give to the common pool by sacrificing (Andreoni, 1995). Subjects generally are given thorough instructions, encouraged to ask questions, and are often given a short quiz to be sure they understand how their choices (combined with choices of others) will determine their money earnings. Economists are also obsessed with offering substantial financial incentives for good performance, and many experiments have been conducted which show that results generalize even when stakes are very large (on the order of several days' or even months' wages).

Since economists are typically interested in whether behavior corresponds to an equilibrium, games are usually played repeatedly to allow learning and equilibration to occur. Because playing repeatedly with the same player can create different equilibria, in most experiments subjects are rematched with a different subject each period in a "stranger" protocol. (In the opposite, "partner" protocol, a pair of subjects know they are playing each other repeatedly.) In a design called "stationary replication", each game is precisely like the one before. This is sometimes called the "Groundhog Day" design, after a movie starring Bill Murray in which Murray's character relives the same day over and over. (At first he is horrified, then he realizes he can learn by trial-and-error because the events of the day are repeated identically.)

After subjects make choices, they are usually given feedback on what the subject they are paired with has done (and sometimes feedback on what all subjects have done), and compute their earnings. Some experiments use the "strategy method" in which players make a choice conditional on every possible realization of a random variable or choice by another player. (For example, in a bargaining game subjects might be asked whether they would accept or reject every offer the other player could make. Their conditional decision is then enacted after the other player's offer is made.) At the end of the experiment, subjects are paid their actual earnings plus a small "show-up" fee

(usually $3-$5). In experimental economics, there is a virtual taboo against deceiving subjects by actively lying about the experimental conditions, such as telling them they are playing another person when they are not (which is quite common in social psychology). A major reason for this taboo is that for successful experimentation subjects have to believe the information that is given to them by the experimenter. In the long run deception can undermine the credibility of the information given to the subjects.

The seven examples we will discuss are prisoners' dilemma (PD), public goods, ultimatum, dictator, trust, gift exchange, and third party punishment games. Table 1 summarizes the definitions of the games (and naturally-occurring examples of them), the predictions of game theory (assuming self-interest and rational play), experimental regularities, and the psychological interpretation of the evidence.

Prisoners' dilemma and public goods games

Table 2 shows payoffs in a typical PD. The rows and columns represent simultaneous choices by two players. Each cell shows the payoffs from a combination of row and column player moves; the first entry is the row player's payoff and the second entry is the column player's payoff. For example, (T,S) in the (Defect, Cooperate) cell means a defecting row player earns T when the column player cooperates, and the column player earns S.

Table 2: Prisoners' dilemma (PD)
(Assumption: T > H > L > S)

	Cooperate (C)	Defect (D)
Cooperate (C)	H, H	S,T
Defect (D)	T, S	L,L

Mutual cooperation provides payoffs of H for each player, which is - by definition of a PD - better than the L payoff from mutual defection. However, if the other player plays C a defector earns the T(emptation) payoff T, which is better than reciprocating and earning only H (since T>H in a PD). A player who cooperates against a defector earns the S(ucker) payoff, which is less than earning L from defecting. Since T>H and L>S, both players prefer to defect whether the other player cooperates or not. So mutual defection is

Table 1: Seven experimental games useful for measuring social preferences

Definition of the Game	Real life Example	Predictions with rational and selfish players	Experimental regularities, References	Interpre
Two players, each of whom can either cooperate or defect. Payoffs are as follows: Cooperate: H,H; Defect: S,T Cooperate: T,S; Defect: L,L H>L, T>H, L>S	Production of negative externalities (pollution, loud noise), exchange without binding contracts, status competition.	Defect	50% choose Cooperate. Communication increases frequency of cooperation Dawes (1980)**	Reciprocate cooper
n players simultaneously decide about their contribution g_i. ($0 \leq g_i \leq y$) where y is players' endowment; each player *i* earns $\pi_i = y - g_i + mG$ where G is the sum of all contributions and $m<1<mn$.	Team compensation, cooperative production in simple societies, overuse of common resources (e.g., water, fishing grounds)	Each player contributes nothing, i.e. $g_i = 0$.	Players contribute 50% of y in the one-shot game. Contributions unravel over time. Majority chooses $g_i=0$ in final period. Communication strongly increases cooperation. Individual punishment opportunities greatly increase contributions. Ledyard (1995)**.	Reciprocate cooper
Division of a fixed sum of money S between a Proposer and a Responder. Proposer offers x. If Responder rejects x both earn zero, if x is accepted the Proposer earns S – x and the Responder earns x.	Monopoly pricing of a perishable good; "11th-hour" settlement offers before a time deadline	Offer $x=\varepsilon$; where ε is the smallest money unit. Any $x>0$ is accepted.	Most offers are between .3 and .5S. $x < .2S$ rejected half the time. Competition among Proposers has a strong x-increasing effect; competition among Responders strongly decreases x. Güth et al (1982)*, Camerer (in press)**	Responder unfair offers recipro
Like the ultimatum game but the Responder cannot reject, i.e., the "Proposer" dictates (S-x, x).	Charitable sharing of a windfall gain (lottery winners giving anony-mously to strangers)	No sharing, i.e., x = 0	On average "Proposers" allocate x=.2S. Strong variations across experiments and across individuals Kahneman et al (1986)*, Camerer (in press)**	Pure alt

Investor has endowment S and makes a transfer y between 0 and S to the Trustee. Trustee receives 3y and can send back any x between 0 and 3y. Investor earns S – y + x. Trustee earns 3y – x.	Sequential exchange without binding contracts (buying from sellers on Ebay)	Trustee repays nothing: x = 0. Investor invests nothing: y = 0.	On average y = .5S and trustees repay slightly less than .5S. x is increasing in y. Berg et al (1995)*, Camerer (in press)**	Trustees sho recipro
"Employer" offers a wage *w* to the "worker" and announces a desired effort level *ê.* If worker rejects (*w*, *ê)* both earn nothing. If worker accepts, he can choose *any e* between 1 and 10. Then employer earns 10e – w and worker earn *w – c(e*). *c(e)* is the effort cost which is strictly increasing in *e*.	Noncontractibility or nonenforceability of the performance (effort, quality of goods) of workers or sellers.	Worker chooses e = 1. Employer pays the minimum wage.	Effort increases with the wage *w*. Employers pay wages that are far above the minimum. Workers accept offers with low wages but respond with e = 1. In contrast to the ultimatum game competition among workers (i.e., Responders) has no impact on wage offers. Fehr et al (1993)*	Workers reci generous wag Employers ap workers' reci offering gene wages.
A and B play a dictator game. C observes how much of amount S is allocated to B. C can punish A but the punishment is also costly for C.	Social disapproval of unacceptable treatment of others (scolding neighbors).	A allocates nothing to B. C never punishes A.	Punishment of A is the higher the less A allocates to B. Fehr and Fischbacher (2001a)*	C sanctions v a sharing nor

ote: ** denotes survey papers, * denotes papers that introduced the respective games.

the only Nash (mutual best-response) equilibrium.[2] This equilibrium is inefficient because mutual cooperation would render both players better off.

Public goods games have a similar incentive structure as PD-games.[3] They can, in fact, be viewed as generalized PDs because the players have incentives to contribute nothing to the public good, but contributions from everyone would make everyone better off. The following experiment illustrates a typical public goods game. There are n subjects in a group and each player has an endowment of y dollars. Each player can contribute between zero and y dollars to a group project. For each dollar that is contributed to the group project *every* group member (including those who contributed nothing) earns $m < 1$ dollars. The return m thus measures the marginal *private* return from a contribution to the group project. Since a subject benefits from the contributions of the others it is possible to free-ride on these contributions. The parameter m also obeys the condition $mn > 1$. The product mn is the total marginal return for the *whole group* from a contribution of one more dollar. For each dollar that is kept by a subject, that subject earns exactly one dollar. The total material payoff π of a subject that contributes g dollars is, therefore, given by $\pi = y - g + mG$ where G is the sum of the contributions of all n group members.[4] Self-interested subjects should contribute nothing to the public good, regardless of how much the other subjects contribute. Why? Because every dollar spent on the group project costs the subject one dollar but yields only a *private* return of $m < 1$. This means, that in equilibrium all self-interested subjects will contribute nothing to the public good. A group of self-interested subjects earns y dollars in this experiment because $G = 0$. But since the total return for the group mn is larger than one, the group as a whole benefits from contributions. If all group members invest their entire endowments y, then $G = ny$ which means each subject earns mny rather than y (which is better because mn is larger than one). Thus, contributing everything to the group project renders all subjects

[2] It is important to note the distinction between outcomes that are measured in field data or paid in experiments, and the utilities or personal valuations attached to those rewards. Game theory allows the possibility that players get utility from something other than their own rewards (e.g., they may feel pride or envy if others earn lots of money). In practice, however, we observe only the payoffs players earn. For the purpose of this chapter, when we assume "self-interest" we mean that players are solely motivated to maximize their own measured earnings in dollars (or food, or some other observed outcome).

[3] There is a huge literature on public good games. For a survey see Ledyard (1995).

[4] In the general case players may have unequal endowments y_i and they may derive unequal benefits m_i from the Public Good G. m_i may also depend non-linearly on G. The material payoff of player i can then be expressed as $\pi_i = y_i - g_i + m_i(G)G$. However, for anthropology experiments it is advisable to keep material payoff functions as simple as possible to prevent that subjects are confused. A particularly simple case is given when the experimenter doubles the sum of contributions G and divides the total 2G among all $n > 2$ group members.

better off relative to the equilibrium of zero contributions, but an individual subject does even better by contributing nothing.

The PD and public goods games are models of situations like pollution of the environment, in which one player's action imposes a harmful "externality" on innocent parties (cooperation corresponds to voluntarily limiting pollution), villagers sharing a depletable resource like river water or fish in a common fishing ground with poor enforcement of property rights (e.g., Ostrom 2000), and production of a public utility like a school or irrigation system that noncontributing "free riders" cannot be easily excluded from sharing. Note also that contributions in public goods games are often in the form of time rather than money—for example, helping to clean up a public park or standing watch for village security. Low rates of voluntary cooperation and contribution in these games might be remedied by institutional arrangements like government taxation (which forces free riders to pay up), or informal mechanisms like ostracism of free riders. (Of course, if ostracism is costly then players should free-ride on the ostracism supplied by others, which creates a second-order public good problem.) Also, when PD and public goods games involve players who are matched together repeatedly, it can be an equilibrium for players to all cooperate until one player defects. Sometimes the experimenter wants to allow for stationary replication but, at the same time wants to prevent the existence of equilibria that involve positive contribution levels. This can be achieved by changing the group composition from period to period such that no player ever meets another player more than once.

In the PD self-interested subjects have an incentive to defect. In the public good game, when $m < 1$, the self-interest hypothesis predicts zero contributions. In experiments, however, subjects in one-period PD-games cooperate about half of the time. In one-period public good games they contribute an average of 40-60 percent of their endowment, but the distribution is typically bimodal with most subjects contributing either everything or nothing. Higher values of the private return m lead to higher contributions. Similar effects are obtained in the PD. An increase in the value of H, relative to T, increases the rate of cooperation. Interestingly, pre-play communication about how much players intend to contribute, which should have no effect in theory, has a very strong positive impact on cooperation levels in both the PD and public good games (Ledyard, 1995; Sally, 1995).

When the public good game is repeated for a finite number of periods interesting dynamic contribution patterns emerge. Irrespective of whether subjects can stay together in the same group or whether the group composition changes from period to period, subjects initially contribute as much as they do in one-period games, but contributions decline

substantially over time. Approximately 60 to 80 percent of all subjects contribute nothing in the final period and the rest contribute little.[5] The first ten periods of Figure 1 show the dynamic pattern of average contributions in a standard public good game like the one described above. Another important fact is that about half the subjects are "conditional cooperators" - they contribute more when others are expected to contribute more and do contribute more (Croson, 1999; Fischbacher, Gächter and Fehr, forthcoming). Conditional cooperation is not compatible with pure self-interest, but consistent with a preference for behaving reciprocally. The studies cited above also indicate that about a third of the subjects are purely self-interested, and never contribute anything.

Why do average contributions decline over time? A plausible explanation is that each group has a mixture of subjects who behave selfishly and others who behave reciprocally. The reciprocal subjects are willing to cooperate if the other group members cooperate as well. However, in the presence of selfish subjects who never contribute, reciprocal subjects gradually notice that they are matched with free-riders and refuse to be taken advantage of. [6]

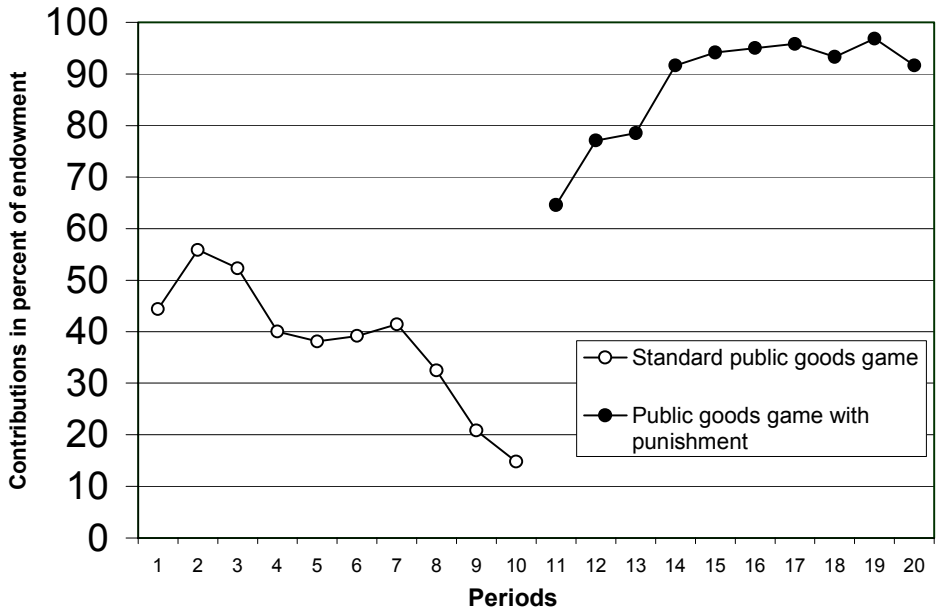
The unraveling of cooperation over time raises the question of whether there are social mechanisms that can prevent the decay of cooperation. A potentially important mechanism is social ostracism. In a series of experiments Fehr and Gächter (2000) introduced a punishment opportunity into the public goods game. In their game there are two stages. Stage one is a public goods game as described above. In stage two, after every player in the group has been informed about the contributions of each group member, each player can assign up to ten punishment points to each of the other players. The assignment of one punishment point reduces the first-stage income of the punished subject by ten percent but it also reduces the income of the punisher. (The punishment is like an angry

[5] Initially, many experimentalists interpreted this as a victory of the self-interest hypothesis (Isaac, McCue and Plott 1985). It was thought that at the beginning of the experiment subjects do not yet fully understand what they rationally should do (even though the incentive to free-ride is usually transparent and is often pointed out very explicitly in the instructions) but over time they learn what to do and in the final period the vast majority of subjects behave self-interestedly. This interpretation is wrong. Andreoni (1988) showed that if one conducts a "surprise" second public good game after the final period of a first game, subjects start the new game with high contribution levels (similar to initial levels in the first game). If players had learned to free ride over time, this "restart" effect would not occur; so the dynamic path that is observed is more likely to be due to learning by conditional cooperators about the presence and behavior of free-riders, rather than simply learning that free-riding is more profitable.

[6] The existence of conditional cooperators may also explain framing effects in public goods and PD games. If, e.g., a PD game is described as the “Wallstreet” game, subjects are likely to have pessimistic expectations about the other players’ cooperation. Conditional cooperators are, therefore, likely to defect in this frame. If, in contrast, the PD is described as a “Community” game, subjects probably have more optimistic expectations about the cooperation of the other player. Hence, the conditional cooperators are more likely to cooperate in this frame.

group member scolding a free-rider, or spreading the word so the free-rider is ostracized--there is some cost to the punisher, but a larger cost to the free-rider.) Note that since punishment is costly for the punisher, the self-interest hypothesis predicts zero punishment. Moreover, since rational players will anticipate this, the self-interest hypothesis predicts no difference in the contribution behavior between the standard public goods game and the game with a punishment opportunity. In both conditions zero contributions are predicted.

Figure 1: Average contributions over time in public good games with a constant group composition (Source: Fehr and Gächter 2000)



The experimental evidence completely rejects this prediction.[7] In contrast to the standard public goods game, where cooperation declines over time and is close to zero in the final period (see the first ten periods in Figure 1), the punishment opportunity causes a sharp jump in cooperation (compare period 10 with period 11 in Figure 1) and a steady increase until almost all subjects contribute their whole endowment. The sharp increase occurs because free-riders often get punished, and the less they give, the more likely punishment is. Cooperators feel that free-riders take unfair advantage of them and, as a consequence,

[7] In the experiments subjects first participated in the standard game for ten periods. After this they were told that a new experiment takes place. In the new experiment, which lasted again ten periods, the punishment opportunity was implemented.

they are willing to punish the free-riders. This induces the punished free-riders to increase cooperation in the following periods. A nice feature of this design is that the actual rate of punishment is very low in the last few periods - the mere threat of punishment, and the memory of its sting from past punishments, is enough to induce potential free-riders to cooperate.

The results in Figure 1 are based on a design in which the same group of players are paired together repeatedly (the "partner" protocol). When the group composition changes randomly from period to period or when subjects are never matched with the same group members again (the "stranger" protocol), cooperation levels are lower than in the partner design, but the dynamic pattern is similar to Figure 1. Interestingly, the punishment pattern is almost the same in the partner and the stranger protocol. This means that, in the partner protocol, the strategic motive of inducing future cooperation is not an important cause of the punishment.

The public goods game with a punishment opportunity can be viewed as the paradigmatic example for the enforcement of a social norm. Social norms often demand that people give up private benefits to achieve some other goal. This raises the question of why most people obey the norm. The evidence above suggests an answer: Some players will punish those who do not obey the norm (at a cost to themselves), which enforces the norm.

Another mechanism that causes strong increases in cooperation is communication (Sally, 1995). If the group members can communicate with each other the unraveling of cooperation frequently does not occur. Communication allows the conditional cooperators to coordinate on the cooperative outcome and it may also create a sense of group identity.

While PD and public goods games capture important component of social life, they cannot typically distinguish between players who are self-interested, and players who would like to reciprocate but believe pessimistically that others will not cooperate or contribute. Three other games have proved useful in separating these two explanations and measuring a wider range of social preferences - ultimatum, dictator, and trust games.

Ultimatum games

Ultimatum games represent a form of take-it-or-leave-it bargaining (Güth, Schmittberger and Schwarze 1982). One player, a Proposer, can make only one proposal regarding the

division of a fixed amount of money *S* between herself and a Responder. The Responder can accept the offer *x*, or reject it, in which case neither player earns anything. If the Responder accepts he earns *x* and the Proposer earns *S* - *x*. In theory, self-interested Responders will accept any positive offer, and Proposers who anticipate this should offer the smallest possible positive amount (denoted by ε in Table 1).

The ultimatum game measures whether Responders will negatively reciprocate, sacrificing their own money to punish a Proposer who has been unfair. In dozens of experiments under different conditions in many different countries, Responders reject offers less than 20 percent of S about half the time. Proposers seem to anticipate this negative reciprocity and offer between 30 and 50 percent of *S*. A typical distribution of offers is given in Figure 2 which shows the data from Hoffman, McCabe and Smith (1994).

Figure 2: Offers and rejections in $10 and $100 ultimatum games (Hoffman, McCabe and Smith, 1996).

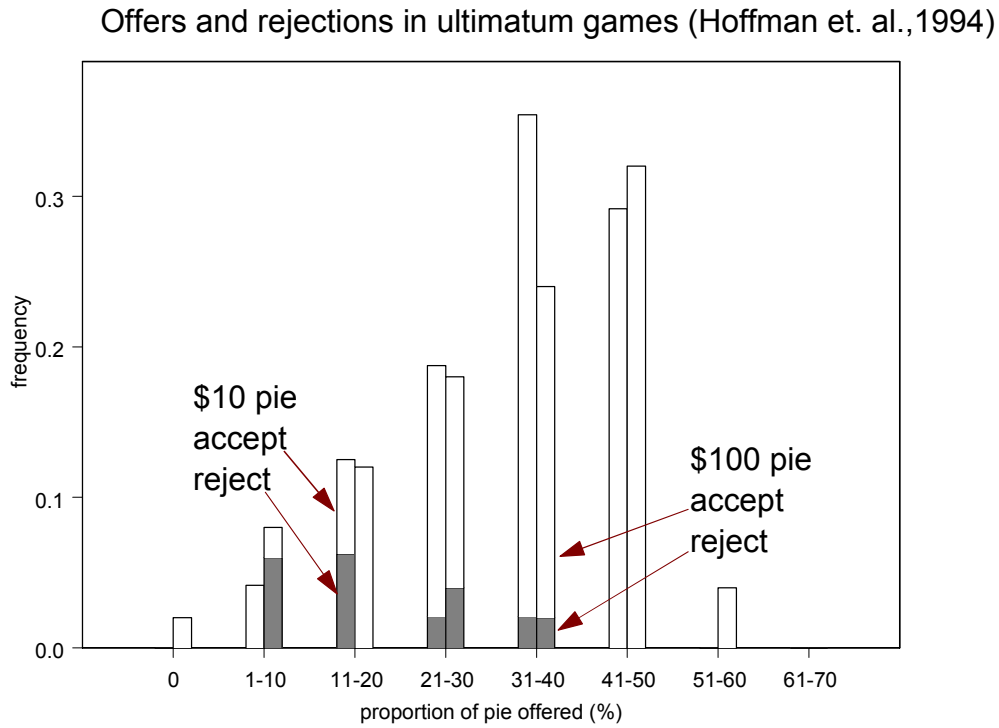
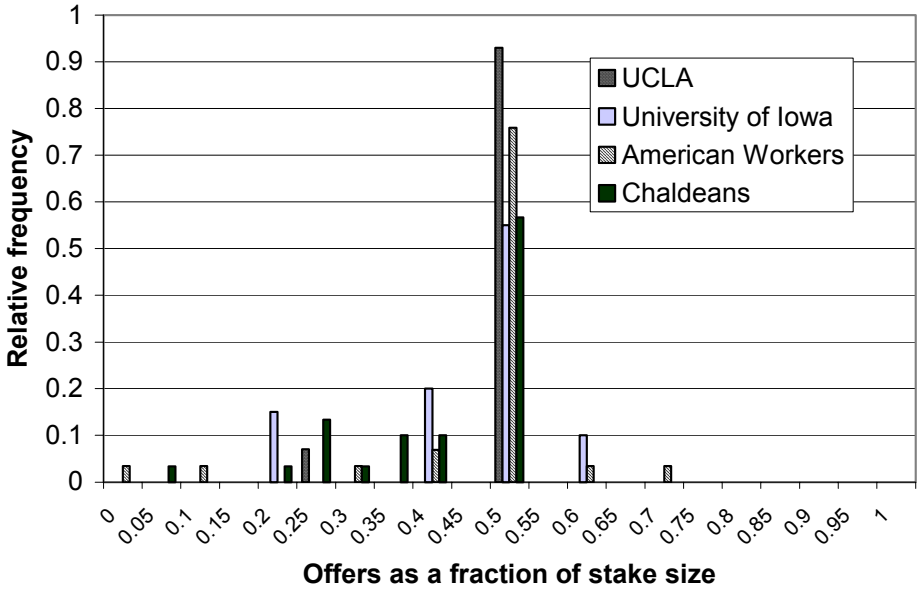


Figure 3 shows offers from experiments with four groups (UCLA graduate students; students from University of Iowa; employees from a large firm in Kansas City (see Burks et al, 2001), and Chaldeans, who are Catholic Iraqis in Detroit (see Smith, 2000)). The offers and rejection rates are generally quite robust across (developed) cultures, levels of stakes (including $100-$400 in the US and 2-3 months' wages in other countries), and changes in experimental methodology (see Camerer, in press). There are weak or unreplicated effects of demographic variables like gender, undergraduate major (economics majors offer and accept less), physical attractiveness (women offer more than half, on average, to more attractive men), and age (young children are more likely to accept low offers). Creating a sense of entitlement by letting the winner of a trivia contest be the Proposer also leads to lower offers and more frequent acceptances.

Figure 3: Distribution of ultimatum offers



An important finding is that competition on the side of the Responders or the Proposers causes large shift in proposals and agreements (Roth et al. 1991; Fischbacher, Fong and Fehr, in preparation). In case of two Responders, e.g., who simultaneously accept or reject the offer x of a single Proposer, the average offer decreases to 20 percent of S. Competition among the Responders induces them to accept less, and Proposers anticipate this and take advantage by offering less. When Proposers compete, by making

simultaneous offers to a single Responder who accepts the single best offer, the average accepted offer rises to 75 percent of *S*.

At a first glance the fact that Responders reject less and Proposers offer more under competitive conditions seems to indicate that the preference for reciprocity is weaker in this situation. But people may have precisely the same kinds of social preferences in two-player and multiplayer games with competition, but act more self-interestedly when there is competition because doing so actually satisfies their preferences. How? Note that a negatively reciprocal Responder is willing to punish a Proposer for an unfair proposal. Under competitive conditions, however, a Responder can only punish the Proposer if the other Responder(s) also reject a given offer. With competition punishment of the Proposer is a public good that is only produced if *all* Responders reject. Since there is always a positive probability to be matched with a self-interested Responder, who accepts every positive offer, the reciprocal Responder's rejection becomes futile. Hence, there is less advantage to rejecting under competition, even if one has a strong preference for reciprocity. Competition essentially makes it impossible for players to express their concern about reciprocity.

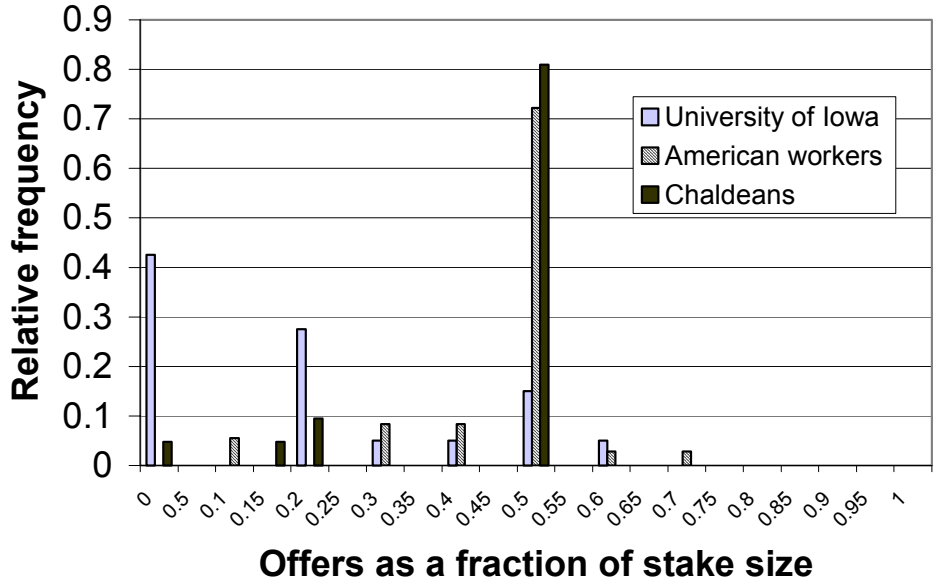
The fact that Proposers offer on average 40 percent of *S* might be due to altruism, a preference for sharing equally, or to a fear that low offers will be rejected ("strategic fairness"). Although rejection rates are lower under competitive conditions there is still a significant rate of rejection. Thus, even under competitive conditions Proposers have reason to fear that very low offers are rejected. Dictator games help separate the fear-of-rejection hypothesis from the other explanations mentioned above because the Responder's ability to reject the offers is removed.

Dictator games

A dictator game is simply a Proposer division of the sum *S* between herself and another player, the Recipient (Kahneman, Knetsch and Thaler 1986; Forsythe et al. 1994). Self-interested proposers should allocate nothing to the Recipient in the dictator game. In experiments with students, Proposers typically dictate allocations that assign the Recipient on average between 10 and 25 percent of *S*, with modal allocations distributed between 50 percent and zero (see Figure 4, from Smith 2000). These allocations are much less than student Proposers offer in ultimatum games, though most players do offer something. Comparing dictator with bilateral ultimatum games shows that fear of rejection is *part* of the explanation for Proposers' generous offers, because they do offer

less when there can be no rejection. But many subjects offer something in the dictator game, so fear of rejection is not the entire explanation. Moreover, the Chaldeans and the employees from Kansas City offer roughly the same in the ultimatum and the dictator game.[8]

Figure 4: Dictator game allocations



The dictator game is a "weak situation" because average allocations can change dramatically with changes in the experimental design. At one extreme, when experimenters take pains to ensure to subjects that their individual decisions cannot be identified by the experimenter (in "double-blind" experiments), self-interest emerges more strongly (among students): About 70 percent of the Proposers allocate nothing and the rest typically allocate only 10-20 percent of *S* (Hoffman et al, 1994). At the opposite extreme, when the eventual recipient of the Proposer's allocation gives a short description of him or herself which the Proposer hears, the average allocation rises to half of S, and allocations become more variable (Bohnet and Frey, 1999). Many Proposers give nothing and others give the entire amount, as if Proposers make an empathetic

[8] Unfortunately, there are so far not many experiments with non-student populations. It is therefore not clear to what extent the results from the Chaldeans (Smith 2000) and from the Kansas City workers (Burks et al., 2001) represent general patterns in non-student populations.

judgment about the recipient's deservingness. These two extremes simply illustrate that dictator allocations can be strongly influenced by many variables (in contrast to ultimatum offers, which do not deviate too far from 30-50% in most previous experiments with students).

Trust and gift exchange games

Dictator games measure pure altruism. An interesting companion game is the "trust game" (Berg, Dickhaut and McCabe 1995). In a trust game an Investor receives an amount of money S from the experimenter, and then can send between zero and S to the Trustee. The experimenter then triples the amount sent, which we term y, so that the Trustee has $3y$. The Trustee is then free to return anything between zero and $3y$ to the Investor. The payoff of the Investor is $S - y + z$ and the payoff of the Trustee is $3y - z$ where z denotes the final transfer from the Trustee to the Investor. The trust game is essentially a dictator game in which the Trustee dictates an allocation, but the amount to be allocated was created by the Investor's initial investment.

In theory, self-interested Trustees will keep everything and repay $z = 0$. Self-interested Investors who anticipate this should transfer nothing, i.e., $y = 0$. In experiments in several developed countries, Investors typically invest about half the maximum on average, although there is substantial variation across subjects. Trustees tend to repay slightly less than y so that trust does not quite pay. The amount Trustees repay increases with y, which can be interpreted as positive reciprocity, or a feeling of obligation to repay more to an Investor who has exhibited trust.

Positive reciprocity like the one that shows up in the trust game has important implications for the enforcement of informal agreements and incomplete contracts. Most social relations are not governed by explicit contracts but by implicit informal agreements. Moreover, when explicit contracts exist they are often highly incomplete, which gives rise to strong incentives to shirk (Williamson, 1985). Economic historians like North (1990) have argued that differences in societies' contract enforcement capabilities are probably a major reason for differences in economic growth and human welfare.

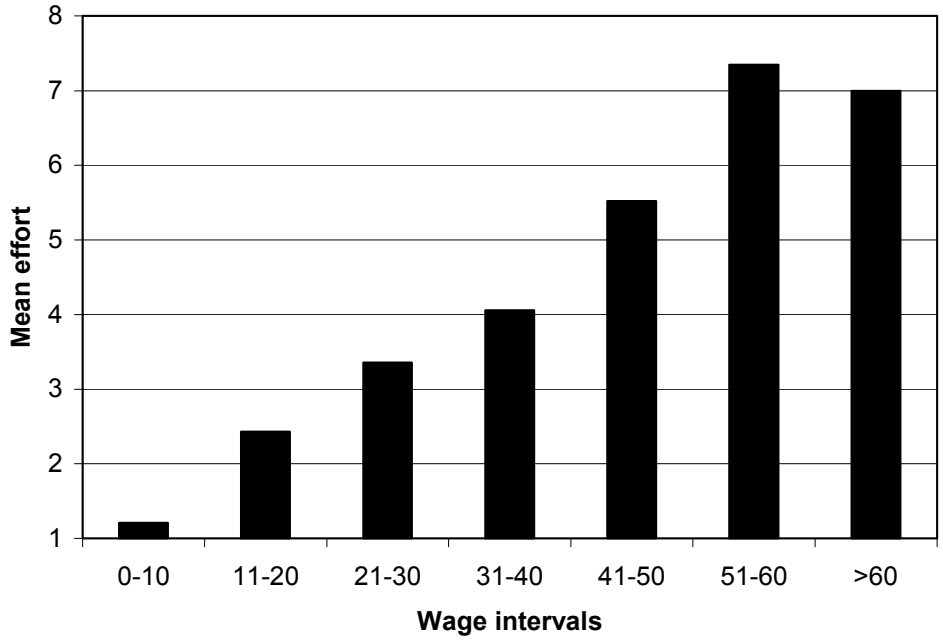
To see the role of reciprocity in the enforcement of contracts, consider the following variant of the gift exchange game (Fehr, Kirchsteiger and Riedl 1993). In the gift exchange game subjects are in the role of employers or buyers and of workers or sellers,

respectively.[9] An employer can offer a wage contract that stipulates a binding wage w and a desired effort level $\hat{e}$. If the worker accepted this offer, the worker is free to choose the actual effort level e between a minimum and a maximum level. The employer always has to pay the offered wage irrespective of the actual effort level. In this experiment effort is represented by a number e between 1 and 10. Higher numbers represent higher effort levels and, hence, a higher profit π for the employer and higher effort costs $c(e)$ for the worker. Thus, the lowest effort level gives the worker the highest material payoff but the highest material payoff for the employer is given at the maximal effort level. Formally, the profit π from the employment of a worker is given by $\pi = 10*e - w$ and the monetary payoff for the experimental worker is $u = w - c(e)$. The crucial point in this experiment is that selfish workers have no incentives to provide effort above the minimum level of $e = 1$ irrespective of the level of wages. Employers who anticipate this behavior will, therefore, offer the smallest possible wage such that the worker just accept the contract offer. Reciprocal workers will, however, honor at least partly generous wage offers with non-minimal, generous, effort choices. The question, therefore, is to what extent employers do appeal to workers' reciprocity by offering generous contracts and to what extent workers honor this generosity.

It turns out that in experiments like this many employers indeed make quite generous offers. On average, the offered contracts stipulate a desired effort of $\hat{e} = 7$ and the offered wage implies that the worker receives 44 percent of the total income that is generated if the worker indeed performs at e = 7. Interestingly, a relative majority of the workers honor this generosity. Most of them do not fully meet the desired effort level but they choose levels above e = 1. A minority of the workers ($\approx$ 30 percent) chooses always the minimal effort. The actual average effort is given by $e = 4.4$ – substantially above the selfish choice of $e = 1$. Moreover, on average there is also a strong positive correlation between effort and wages indicating positive reciprocity. A typical effort-wage relation is depicted in Figure 5. Thus, although shirking exists in this situation the evidence suggests that in response to generous offers, a relative majority of the people are willing to put forward extra effort above what is implied by purely pecuniary considerations.

[9] In the following we stick to the employer-worker framing although the experiment could also be presented in a buyer-seller frame. The gift-exchange experiment has been conducted in both frames with virtually the same results.

Figure 5: Effort-wage relation in the gift exchange game
(Source: Fehr, Gächter and Kirchsteiger 1997)



Similar to the ultimatum game the regularities in the gift exchange game are quite robust with regard to stake levels. In experiments in which subjects earned on average between two and three times their monthly incomes the same wage and effort patterns prevail. Another important result is obtained if there is competition between the workers – similar to the Responder competition in the ultimatum game. While in the ultimatum game with Responder competition Proposers make much lower offers compared to the bilateral case, competition has *no* impact on wages in the gift exchange game. The reason for this striking result is that it does not pay for employers to push down wages because reciprocal workers respond to lower wages with lower effort levels.

Third party punishment games

Many small scale societies are characterized by extensive food-sharing. A simple game to examine whether food sharing is a social norm that is enforced by social sanctions has been conducted by Fehr and Fischbacher (2001a). The game is called "third party

punishment game" and has three players. The game between player A and player B is just a dictator game. Player A receives an endowment of 100 tokens of which he can transfer any amount to player B, the Recipient. Player B has no endowment and no choice to make. Player C has an endowment of 50 tokens and observes the transfer of player A. After this player C can assign punishment points to player A. For each punishment point assigned to player A player C has costs of 1 token and player A has costs of 3 token. Since punishment is costly a self-interested player C will never punish. However, if there is a sharing norm player C may well punish player A if A gives too little.

In fact, in the above experiments player As are never punished if they transferred 50 or more tokens to player B. If they transferred less than 50 tokens the punishment was the stronger the less player A transferred. In case that player A transferred nothing she received on average 9 punishment points from player C, i.e. the payoff of player A was reduced by 27 tokens. This means that in this three-person game it was still beneficial, from a selfish point of view, for player A to give nothing compared to an equal split, say. If there is more than one player C, who can punish player A, this may, however, no longer be the case.

Another interesting question is to what extent cooperation norms are sustained trough the punishment of free-riders by **third parties**. We have already seen that in the public goods game with punishment strikingly high cooperation rates can be enforced through punishment. In this game each contribution to the public good increases the payoff of each group member by 0.4. Thus, if a group member free-rides instead of cooperation she directly reduces the other group members' payoff. In real life there are, however, many situations in which free-riding has a very low, indeed almost imperceptible, impact on the payoff of particular other individuals. The question then is, whether these individuals nevertheless help enforcing a social norm of cooperation. In case they do a society greatly magnifies its capability of enforcing social norms because every member of a society acts as a potential policemen.

It is relatively ease to construct cooperation games with punishment opportunities for third (unaffected) parties. Fehr and Fischbacher (2001a), e.g., have conducted PDs in which a member of the two-person group, who played the PD, observes a member of some other group, who also played the PD. Then the member of the first group can punish the member of the second group. Thus, each member could punish and could be punished by somebody outside the own two-person group. It was ensured that reciprocal punishment was not possible, i.e. if subject A could punish subject B, subject B could not punish A but only some third subject C. It turns out that the punishment by third parties is

surprisingly strong. It is only slightly weaker than second party (within group) punishment.

3. Theories of social preferences

Within economics, the leading explanation for the patterns of results described above is that agents have social preferences (or "social utility") which take into account the payoffs and perhaps intentions of others. Roughly speaking, social preference theories assume that people have stable preferences for how money is allocated (which may depend on who the other player is, or how the allocation came about), much as they are assumed in economics to have preferences for food, the present versus the future, how close their house is to work, and so forth.[10]

Cultural anthropologists and evolutionary psychologists have sought to explain the origin of these preferences. One idea is that in the environment of evolutionary adaptation (EEA) or ancestral past, people mostly engaged in repeated games with people they knew. Evolution created specialized cognitive heuristics for playing repeated games efficiently. It is well-known in game theory that behavior which is optimal for a self-interested actor in a one-period game with a stranger - such as defecting or free riding, accepting all ultimatum offers - is not always optimal in repeated games with partners. In a repeated ultimatum game, for example, it pays to reject offers to build up a reputation for being hard to push around, which leads to more generous offers in the future. In the unnatural habitat view, subjects cannot "turn off" the habitual behavior shaped by repeated-game life in the EEA when they play single games with strangers in the lab. An important modification of this view is that evolution did not equip all people with identical hard-wired instincts for playing games, but instead created the capacity for learning social norms. The latter view can explain why different cultures would have different norms.

[10] A different interpretation is that people have rules they obey about what to do—such as, share money equally if you haven't earned it (which leads to equal-split offers in the ultimatum game) (Güth 1995). A problem with the rule-based approach is that subjects **do** change their behavior in response to changes in payoffs, in predictable ways. For example, when the incremental payoff from defecting against a cooperator (denoted T - H above) is higher, people defect more often. When a player's benefit *m* of the public good is higher, they contribute more. When the social return from investing in a trust game is lower, they invest less. Any rule-based account must explain why the rules are bent by incentives, and such a theory will probably end up looking like a theory of social preferences which explicitly weighs self-interest against other dimensions.

As is common in evolutionary explanations, the unnatural habitat theory assumes the *absence* of a module or cognitive heuristic which could have evolved but did not - the capacity to distinguish temporary one-shot play from repeated play. If subjects had this ability they would behave appropriately in the one-shot game. In principle it is testable whether people have the ability to distinguish temporary one-shot play from repeated play. Fehr and Fischbacher (2001b) did this in the context of the ultimatum game.

They conducted a series of ten ultimatum games in two different conditions. In both conditions subjects played against a different opponent in each of the ten iterations of the game. In each iteration of the baseline condition the Proposers knew nothing about the past behavior of their current Responders. Thus, the Responders could not build up a reputation for being "tough" in this condition. In contrast, in the reputation condition the Proposers knew the full history of the behavior of their current Responders, i.e., the Responders could build up a reputation for being "tough". In the reputation condition a reputation for rejecting low offers is, of course, valuable because it increases the likelihood to receive high offers from the Proposers in future periods.

If the Responders understand that there is a pecuniary payoff from rejecting low offers in the reputation condition one should observe higher acceptance thresholds in this condition. This is the prediction of the social preferences approach that assumes that subjects derive utility from both their own pecuniary payoff and a fair payoff distribution. If, in contrast, subjects do not understand the logic of reputation formation and apply the same habits or cognitive heuristics to both conditions one should observe no systematic differences in Responder behavior across conditions. Since the subjects participated in both conditions it was possible to observe behavioral changes at the individual level. It turns out that the vast majority (slightly more than 80 percent) of the Responders increase their acceptance thresholds in the reputation condition relative to the baseline condition. This contradicts the hypothesis that subjects do not understand the difference between one-shot and repeated play.

The above experiment informs us about the proximate mechanisms that drive Responder behavior in the ultimatum game. Whatever the exact proximate mechanisms will turn out to be, a hypothesis that is based on the story that subjects do not really understand the difference between one-shot and repeated play seems to be wrong. A plausible alternative hypothesis is that Responders face strong emotions when faced with a low offer and that these emotions trigger the rejections. These emotions may be the result of repeated game interactions in our ancestral past and may not be fine-tuned to one-shot interactions. For modeling purposes, behaviorally relevant emotions can be

captured by appropriate formulations of the utility function. This is exactly what theories of social preferences do.

The challenge for all the social preference theories (and evolutionary explanations of their origins) is to explain a lot of results in different games with one model, and make new predictions which survive attempts at falsification. For example, why players contribute in the standard public goods games at first, then stop contributing; why they punish and contribute in the public goods game with punishment opportunities; why Responders reject unfair offers; why Proposers in the dictator game give away money; why many Trustees repay trust; why third parties punish defection in the PD and unfair allocations in the dictator game and why competition causes more unequal divisions in ultimatum games but has no impact in gift exchange games.

Two flavors of models have been proposed—models of inequality-aversion and models of reciprocity. In inequality-aversion theories, players prefer more money and also prefer that allocations be more equal. Attempting to balance these two goals, players will sacrifice some money to make outcomes more equal. For example, in the theory of Fehr and Schmidt (1999) the players' goals are formalized as follows. Let x_i denote the material payoff of player i and x_j the material payoff of player j. Then the utility of player i in a two player game is given by $U_i(x) = x_i - \alpha_i(x_j - x_i)$ if player i is worse off than player j ($x_j - x_i \geq 0$), and $U_i(x) = x_i - \beta_i(x_i - x_j)$ if player i is better off than player j ($x_i - x_j \geq 0$). α_i is a constant that measures how much player i dislikes disadvantageous inequality while β_i measures how much i dislikes advantageous inequality. When α_i and β_i are zero player i is selfish. Fehr and Schmidt also assume that, in general, players dislike advantageous inequality less than disadvantageous inequality, i.e., $0 \leq \beta_i \leq \alpha_i$ and $\beta_i < 1$. For α_i they assumed no upper bound.[11]

A further important ingredient of this theory is that the population of players is assumed to be heterogeneous. In particular, it is assumed that there is a substantial fraction of purely selfish players that coexists with inequity averse players. This model predicts all the regularities mentioned above: Small offers in the ultimatum game are rejected by players with a positive α ("envy") and positive allocations in dictator games occur when players have a positive β ("guilt"). A positive β also explains why Trustees repay some money to Investors in the trust game and why players who expect that the

[11] In the general n-person case the utility function of Fehr and Schmidt is given by $U_i(x) = x_i - \alpha_i \frac{1}{n-1}\sum_{j \neq i} \max\{x_j - x_i, 0\} - \beta_i \frac{1}{n-1}\sum_{j \neq i} \max\{x_i - x_j, 0\}$. The term $\max\{x_j - x_i, 0\}$ denotes the maximum of $x_j - x_i$ and 0. It measures the extent to which there is disadvantageous inequality between player i and j.

other player(s) cooperate in PD and public goods games reciprocate cooperation rather than defecting or free-riding. The theory is consistent with the fact that in the ultimatum game with responder competition the responders reject much less than in the bilateral ultimatum game and why in the gift exchange game responder competition does not matter. It also is consistent with (third party) punishment in the PD, the dictator game and the public goods game. For a quick illustration, consider the PD in Table 3. Note that the numbers in Table 3 represent material payoffs and not utilities.

Table 3: Representation of Prisoners' dilemma (PD) in terms of material payoffs

	Cooperate (C)	Defect (D)
Cooperate (C)	2, 2	0, 3
Defect (D)	3, 0	1, 1

Table 4: Utility representation of PD in Table 3

	Cooperate (C)	Defect (D)
Cooperate (C)	2, 2	$0 - 3\alpha$, $3 - 3\beta$
Defect (D)	$3 - 3\beta$, $0 - 3\alpha$	1, 1

In Table 4 we show the utilities that are attached with the material payoffs of Table 3 if both players have identical preferences with $\alpha > 0$ and $\beta > 0$. In Fehr and Schmidt's theory, if player 2 (the column player) is expected to cooperate, player 1 (the row player) faces a choice between material payoff allocations (2,2) and (3,0). The social utility of (2,2) is $U_1(2,2) = 2$ because there is no inequality. The social utility of (3,0), however, is $U_1(3,0) = 3 - 3\beta$ because there is inequality that favors the row player. Therefore, player 1 will reciprocate the expected cooperation of player 2 **if** $\beta > 1/3$[12] (i.e., if player 1 feels sufficiently "guilty" from defecting). If player 1 defects and player 2 cooperates the payoff of player 2 is $U_2(3,0) = 0 - 3\alpha$; if player 2 defected instead the utility would be 1. This means that player 2 will **always** reciprocate defection because cooperating against a

[12] Note that if the temptation payoff is raised from 3 to T, then a player cooperates if $\beta > (T-2)/T$. Since the latter expression converges to 1 as T grows larger, a player with a fixed β who cooperates at a T near 2 will switch to defection at some point as T grows large; so the model predicts the correct (empirically observed) response to the change in payoff structure.

defector yields less money **and** more envy.[13] Table 4 shows that if $\beta > 1/3$, there are two (mutual best response) equilibria: (cooperate, cooperate) and (defect, defect). In utility terms, inequality averse players no longer face a PD. Instead, they face a coordination or assurance game with one efficient and one inefficient equilibrium (the same as the "stag hunt" game described below). If the players believe that the other player cooperates, it is rational for each of them to cooperate, too.

Inequity averse players are thus conditional cooperators. They cooperate in response to (expected) cooperation and defect in response to (expected) defection. The theory is, therefore, also consistent with framing effects in the PD (and in public goods games). If the framing of the game makes, e.g., the players more optimistic about the other players' cooperation, the inequity averse players will cooperate more.

Inequality-aversion theories are simplified because they include only the other players' material payoffs into the calculation of social utility. Reciprocity theories include other players' actions and, in particular, the intention behind the action, as well. In one important formal reciprocity theory (Rabin, 1993), player A forms a judgment about whether another player B has sacrificed to benefit (or harm) her. A likes to reciprocate, repaying kindness with kindness, and meanness with vengeance.

In the PD Table 3, for example, suppose the row player is planning to cooperate. Then the column player's choice essentially determines what the row player will get. Since row's possible payoffs are 2 and 0, let's take the average of these, 1, to be a "fair" payoff. By choosing to cooperate, the column player "awards" the row player the payoff of 2, which is "nice" because it's greater than the fair payoff of 1.[14] Rabin proposes a utility function in which niceness has a positive value and meanness has a negative value, and players care about their own dollar payoffs and the **product** of their own niceness and the niceness of the other player. Thus, if the other player is nice (positive niceness) they want to be nice too, so the product of nicenesses will be positive. But if the other player is mean (negative niceness) they want to be negative too so the product of nicenesses will be positive. While Rabin's theory is more analytically difficult than other theories, it captures the fact that a single player may behave nicely or meanly depending on how they expect to be treated - it locates social preferences and emotions in the combination of a person, their partner, and a game, rather than as a fixed personal attribute.

[13] This also means that if a selfish and an inequity averse player are matched, and the inequity averse player knows that the other player is selfish, the unique equilibrium is (defect, defect). The reason is that the inequity averse player knows that the other player will defect and hence she will defect, too.

[14] The degree of niceness is formalized by taking the difference between the awarded and fair payoffs, normalized by the range of possible payoffs. In this example, niceness is (2-1)/(2-0)=1/2.

There are also hybrid models that combine the notions of reciprocity with models of social preferences based on own and other players' material payoffs. Charness and Rabin (2000), e.g., proposed a hybrid model in which players care about their own payoffs, and about a weighted average of the lowest payoff anybody receives (a "Rawlsian" component) and the sum of all payoffs (a "utilitarian" component). Their theory has a hidden aversion to inequality through the emphasis on the lowest payoff. In addition, players also care about the actions of the others. Falk and Fischbacher (1999) proposed a model that combines reciprocity and inequality aversion. Both the model of Charness and Rabin and of Falk and Fischbacher explain some data that Fehr-Schmidt's theory cannot explain. This increase in explanatory power comes, however, at a cost because these models are considerably more complicated.

There are an increasing number of experiments that compare predictions of competing theories. One important result of these experiments is that there is evidence for reciprocity beyond inequality-aversion. Players do not only care about the allocation of material payoffs. They also care about the actions and the intentions of the other players.

Regardless of which models are most accurate, psychologically plausible, and technically useful, the important point for social scientists is that a menu of games can be used to measure social preferences, like the extent to which people weigh their monetary self-interest with the desire to reciprocate (or limit inequality), both negatively (in ultimatum games) and positively (in trust games), and with pure altruism (in dictator games). Dozens of experiments in many developed countries, with a wide range of instructions, subjects, and levels of stakes, have shown much regularity. And simple formal theories have been proposed which can account for findings that appear to be contradictory at first blush (e.g., sacrificing money to harm somebody in an ultimatum game, and sacrificing to help somebody in PD or trust games). Exploring behavior in these games in a much wider range of cultures, at various stages of economic development and with varying patterns of sharing norms, governance structures, and so forth, will undoubtedly prove interesting and important. In addition, anthropological studies in remote field sites will serve as an important empirical reminder for economists and psychologists who currently study these games about how very narrow the range of cultures they study is.

4. Why do game experiments? And which games?

A central advantage of experimental games is comparability across subject pools (provided great care is taken in controlling for differences in language, purchasing power of outcomes, interactions with experimenters, and so forth). While comparability is clearly not perfect, it is surely as good as most qualitative measures. A further advantage is replicability. The fact that experiments are replicable is a powerful tool for creating consensus about the fact and their interpretation in the scientific community

In fact, experiments conducted in the field by anthropologists may actually have two large advantages compared to lab experiments in Western countries which usually (though not always) use college students as experimental subjects. First, since anthropologists are in the field for long periods of time, the cost of collecting data is rather low. (Most contributors to this volume often noted that the experiment was unusually fun for participants, probably more so than for college students raised in a world of Nintendo, 500-channel cable TV, and web surfing.) Second, the amount of funds budgeted by granting agencies in developed countries for subject payments typically have extraordinary purchasing power in primitive societies. As a result, it is easy for anthropologists to test whether people behave differently for very large stakes, such as a week or month of wages, compared to low stakes. Such comparisons are important for generalizing to high-stakes economic activity, but are often prohibitively expensive in developed countries.

Games impose a clear structure on concepts which are often vague or fuzzy. Social scientists often rely on data like the General Social Survey, in which participants answer questions such as, “In general, how much do you trust people?” on a 7-point Likert scale. It would be useful to have questions about trust which are more concrete, tied to actual behavior, and likely to be interpreted consistently across people. A question like “How much of $10 would you place in an envelope, knowing it will be tripled and an anonymous person will be keep as much as they like and give the rest back to you?" is arguably a better survey question-- it is more concrete, behavioral, and easy to interpret. Note that anthropologists also study their subjects much more carefully than experimental psychologists and economists do, so they often have lots of behavioral data to correlate with behavior from experimental games.

Of course, games are reductions of social phenomena to something extremely simple, but they can always be made more complex. A painter who first sketches a line drawing on a blank canvas has reduced a complex image to two dimensions of space and

color. But the line drawing reduction is also a platform on which more complex images can be restored (e.g., it can be painted over to give the dimension of color and the illusion of depth).

From a technical point of view it is often useful to apply the so-called strategy method in experiments. In the ultimatum game, e.g., a strategy for the Responder stipulates a Yes or No response *for each possible offer*. A simple way of eliciting a Responder strategy is the elicitation of the Responder's minimal acceptable offer, x^{min}. If the actual offer is below x^{min}, it is rejected, if it is above x^{min}, it is accepted. This method has the big advantage that the experimenter not only knows the Responder's response to the actual offer but also to all other feasible offers. Very often most offers in the ultimatum game are close to the equal split so that there are no rejections. In this case the experimenter learns little about the willingness to accept or reject low offers unless the strategy method is applied.

In simple societies the strategy method may sometimes be too complicated for the subjects. In this case it is advisable to restrict the set of feasible offers. For example, in the ultimatum game the experimenter may only allow a 90:10 offer and a 50:50 offer, and the Responder then has to indicate his response to both potential offers before he knows the actual offer. For similar reasons as in the ultimatum game, the strategy method, is of course, also useful in many other games like, e.g., the trust or the third party punishment game. Knowing the Trustee's response to all feasible investments in the trust game, or player C's punishment of player A for all feasible transfers player A can make to player B in the third party punishment game, provides a lot more information compared to the usual method.

The experimental games described in this chapter are line drawings, to which richness can be added. For example, most of the games we described are only played once without communication (the soundtrack of life is muted) and without mutual identification of who the other players are (like the Magritte painting "The Lovers" in which two people kiss with their heads shrouded in cloth). Conducting experiments this way is obviously **not** a deliberate choice to model a world in which people don't talk and only meet hooded strangers (although it might be appropriate for nearly-anonymous internet transactions). Instead, this baseline design is a stark control condition which can be used to study the effect of communication, by comparing results in the control condition with experiments in which communication is allowed (turning up the soundtrack volume) and mutual identification is allowed (removing the hoods).

Other games social scientists might find useful

While the games described above have been studied most widely (including by anthropologists; see this volume) other games or treatments might also be of interest. This section describes four of them.

Measuring moral authority in dictator games

As noted above, the dictator game is a weak situation in the sense that a wide variety of treatment variables—instructions, entitlement, experimental control for "blindness" to individual allocations, identification of recipients, etc.—affect allocations significantly. The fact that preferences are malleable suggests a way to measure moral authority, which was very cleverly suggested by Caroline Lesogoral (Jean Ensminger's student). Collect a group of subjects. Have a person A suggest a way the subjects should play the dictator game. Then have the subjects play. The extent to which subjects adhere to A's recommendation is a measure of A's moral authority or ability to create norms which are adhered to.

Coordination: Assurance and threshold public goods games

Table 5 shows a game called "stag hunt", also known as an "assurance game" or Wolf's Dilemma. The game is identical to the PD in structure except for one crucial difference: It is better to reciprocate cooperation, because the material payoff to defecting when the other player cooperates is lower than the material payoff from cooperating. If there are strong synergies or "complementarities" from the cooperative choices of two players, or if free riders are punished after they defect, then the PD game is transformed into stag hunt.[15]

[15] Recall that when players are inequality averse the PD, when represented in social utility terms, is transformed into an assurance game. From an experimental viewpoint, this is, however, different from an assurance game where the payoffs are monetary. While the experimenter has full control over the monetary payoffs we can never be sure about the preferences of the players.

The game is called stag hunt after a story in Jean-Jacques Rousseau about hunters who can choose to hunt a large stag with others, which yields a large payoff if everyone else helps hunt the stag, or can hunt for rabbit on their own. An example familiar to anthropologists is hunting for large animals like whales (see Alvard, this volume), in which the marginal hunter's presence can be crucial for a successful hunt. Stag hunt is a "coordination game" because there is more than one Nash equilibrium, and players would like to find a way to coordinate their choices on one equilibrium rather than mismatch. Since stag is a best-response to hunting stag, (stag, stag) is an equilibrium; but so is (rabbit, rabbit).

Table 5: The "stag hunt" or assurance game

	stag	rabbit
stag	2, 2	0, 1.5
rabbit	1.5, 0	1, 1

Stag hunt is closely related to "threshold" public goods games (also called the "volunteer's dilemma"). In these games there is a threshold of total contribution required to produce the public good. If n-1 players have contributed, then it pays for the n^{th} player to pitch in and contribute, since her share of the public good outweighs the cost of her marginal contribution.

The central feature of the PD is whether the other player has social preferences that induce her to cooperate (acting against her self-interest) *and* whether the player himself gets social utility from reciprocating cooperation. Stag hunt is different: Because players get a higher *material* payoff from reciprocating the cooperative choice (stag), all they need is sufficient assurance that others will hunt stag (i.e., a probability of playing stag above 2/3, which makes the expected payoff from stag higher than the expected payoff from rabbit) to trigger their own stag choice. PD is about cooperativeness; how cooperative is player 1 *and* how cooperative does he expect player 2 to be. Stag hunt is solely about perceptions of whether others are likely to cooperate. Experiments with coordination games like stag hunt show that, perhaps surprisingly, the efficient (stag,stag) outcome is not always reached. Pre-play communication helps. Social structure has an interesting effect: If a population of players are matched randomly each period, the tendency to play stag is higher than if players are arrayed on a (virtual) circle and play only their neighbors each period.

Stag hunt could be useful to measure whether a culture has a norm of playing "stag" when the cooperative action is risky.

Status in bargaining

Table 6 shows a game called "battle of the sexes" (BoS). In this game two players simultaneously choose a strategy we have labeled R and C. If the players mismatch they get nothing. If they match on R the row player gets the higher payoff of 3 and the column player gets 1. The payoffs are opposite if they match on C. The game is called "battle of the sexes" after a hoary story about a husband and wife who would like to attend an event together, but the husband prefers boxing while the wife prefers ballet.

BoS is a classic "mixed-motive" game because the players prefer to agree on something than to disagree, but they disagree on what to agree on. Alternatively, think of the game as a bargaining game in which the players will split 4 if they can agree how to split it (but it must be uneven, 3:1 or 1:3) and earn nothing otherwise.

Table 6: Battle of the sexes game (BoS)

	R	C
R	3, 1	0, 0
C	0, 0	1, 3

In experiments with payoffs like Table 6, players tend to choose their preferred strategy (row chooses R, column chooses C) around 65% of the time, which means they mismatch more than half the time (see Camerer, in press, chapter 7). Since mismatches yield nothing, the game cries out for some social convention or coordinating device which tells players which one of them gets the larger payoff; in principle, the player who gets less should go along with the convention since getting 1 is better than mismatching and getting nothing.

Any commonly-understood variable which produces consistent matches in a pair of players can be interpreted as an indicator of **status**. A wonderful illustration of this is Holm's (2000) experiments on BoS and gender. He ran experiments in which men and women played BoS games (simultaneously, with no communication) with players of the

same sex and opposite sex. Take the row player's view. When women played with men, the women (in the row player position) were more likely to play C and men (in the row player position) where more likely to play R, compared to when they played with subjects of the same gender. The players played as if they all respected a social convention in which women get the smaller share and the men the larger share. Remarkably, women actually earned a larger average payoff playing against men than playing against other women! The reason for this is that earning 1 with a high probability is better than trying to earn 3 but mismatching very frequently.

We interpret these results as evidence that males have status. An agreed-upon status variable has two interesting effects in these games: It increases collective gains (by minimizing mismatches); and it creates greater wealth for the high-status group than for the low-status group. The latter effect, of course, can spark a self-fulfilling spiral in which, if wealth itself creates status, the rich get status and get richer too.[16]

Since concepts of hierarchy, privilege, and status are central in anthropology (and in sociology), games like BoS which reveal status relations (and show their economic impact) could prove useful. Game-theoretic revelation of status also provides a way for economists to comprehend such concepts, which do not fit neatly into primitive economic categories like preferences and beliefs.

Shared understanding and cultural homogeneity in matching games

In 1960 Schelling drew attention to simple "matching games", in which players choose an object from some category, and earn a fixed prize if their objects match. For example, subjects who are asked to choose a place and time to meet in New York City often choose noon at Grand Central Station, or other prominent landmarks like Central Park or the Statue of Liberty. Careful experiments by Mehta, Starmer and Sugden (1994) show the same effect. Asked to name a mountain, 89% of subjects picked Mt. Everest; naming a gender, 67% picked "man"; naming a relative, 32% picked "mother" (20% picked "father"); asked to pick a meeting place in London, 38% picked Trafalgar Square; and so forth.

[16] An alternative interpretation is that Bos-play reflects to which extent the aggressiveness of the other player is common knowledge. If all women believe that men are more aggressive, it pays for them to give in. Yet, if wealth creates status, than the greater aggressiveness of men ultimately also confers status.

From a game-theoretic viewpoint, matching games with a large choice set have lots and lots of equilibria. Schelling's point was that shared world knowledge often picks out a psychologically prominent "focal" point. A focal point is the right choice if "everybody knows" it's the right choice. The extent of shared understanding can be measured by how well subjects match. We suggest this as a measure of cultural homogeneity. For example, Los Angeles is a diverse patchwork of local communities of wildly varying ethnicity. Asked to choose a meeting place in LA (playing the game with their own ethnic or geographical community), Koreans might choose the corner of Western and Wilshire (the heart of "Koreatown"), those from south beach might choose "The Strand" (a boardwalk by the ocean), West Hollywood gays might choose local bar Mickey's, Hollywood Hills trendies would choose Skybar, and so forth. The fact that most readers haven't heard of all these "famous" places is precisely the point. The degree to which a group coordinates on a culturally-understood meeting place seems like a good measure of overall cultural homogeneity. (If they don't agree, they aren't a group-- at least not a group with shared cultural knowledge.)

Camerer and Weber (2001) use matching games, with a linguistic twist, to study endogenous development of culture and cultural conflict. In their experiments, a pair of subjects are each shown 16 pictures which are very similar (e.g., scenes of workers in an office). One subject is told that eight of the pictures have been selected as targets. This subject, the director, must describe the pictures to the second subject, so that the second subject chooses the correct pictures as quickly as possible. (They earn money for accuracy and speed.) Since the subjects have never seen these pictures before, they must create a homemade language to label the pictures. Because they are under time pressure, with repeated trials they create a very pithy "jargon" to describe the distinctive features of a picture as briefly as possible. Their homemade language is one facet of culture (albeit designed to accomplish a specific purpose-- commonly-understood labeling of novel objects). Cultural conflict can be studied by combining two separate groups, whose jargon tend to be different.

These paradigms can be used to measure or create shared understanding, with economic incentives to reveal shared understanding or create it quickly. These could prove useful in anthropology too for measuring cultural homogeneity and dimensions of shared perception.

5. Conclusions

Game theory has proved useful in a wide range of social sciences in two ways: By providing a taxonomy of social situations which parse the social world; and by making precise predictions about how self-interested players will actually play. Behavior in experiments which carefully control players' strategies, information, and possible payoffs shows that actual choices often deviate systematically from the game-theoretic prediction based on self-interest. These deviations are naturally interpreted as evidence of social norms (what players expect and feel obliged to do) and social preferences (how players feel when others earn more or less money). This evidence is now being used actively by economists to craft a parsimonious theory of social preferences which can be used to explain data from many different games in a simple way that makes fresh predictions. Since anthropologists are often interested in how social norms and preferences emerge, evolve, and vary across cultures, these games could provide a powerful tool for doing empirical anthropology. In addition to measuring social preferences and social norms experimental games may also be used for measuring moral authority, players beliefs about other players' actions in coordination games, cultural homogeneity and status effects in bargaining.

References

Andreoni, James (1995); “Warm-Glow versus Cold-Prickle: the Effects of Positive and Negative Framing on cooperation in Experiments”, *Quarterly Journal of Economics* 110, 1-22.

Berg, Joyce, John Dickhaut, and Kevin McCabe (1995); "Trust, Reciprocity and Social History," *Games and Economic Behavior* X, 122-142.

Bohnet Iris and Bruno S. Frey (1999). Social Distance and Other-regarding Behavior in Dictator Games: Comment. American Economic Review, 89 (March), 335-339.

Burks, Stephen, Gary Carpenter, Jeffrey Carpenter and Eric Verhoegen (in preparation); “High Stakes Bargaining with Non-Students”, 2001.

Camerer, Colin F. *Behavioral game theory: Experiments on strategic interaction.* Princeton: Princeton University Press, in press.

Camerer, Colin F. and Weber, Roberto (in preparation); “Cultural conflict: An experimental approach”, 2000.

Camerer and Ho (1999); Experience-weighted attraction (EWA) learning in normal-form games," Econometrica 67, 827-874.

Charness, Gary, and Rabin, Matthew (2000). “Social Preferences: Some Simple Tests and a New Model.” *Mimeo*, University of California at Berkeley.

Croson, Rachel T. A. (1999); " Theories of Altruism and Reciprocity: Evidence from Linear Public Goods Games," Discussion Paper, Wharton School, University of Pennsylvania.

Davis, Douglas and Holt, Charles (1993); Experimental Economics, Princeton University Press, Princeton, New Jersey.

Dawes, Robyn M. (1980); Social Dilemmas, *Annual Review of Psychology* 31, 169-193.

Falk, Armin and Urs Fischbacher (1999); “A Theory of Reciprocity.” Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 6.

Fehr, Ernst and Urs Fischbacher (2001a); “Third Party Punishment”, mimeo, University of Zürich.

Fehr, Ernst and Urs Fischbacher (2001b); “Retaliation and Reputation”, mimeo, University of Zürich.

Fehr, Ernst, Georg Kirchsteiger, and Arno Riedl (1993); „Does Fairness prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics* CVIII, 437-460.

Fehr, Ernst and Klaus M. Schmidt, 1999. "A Theory of Fairness, Competition and Co-operation." *Quarterly Journal of Economics* 114, 817-868.

Fehr, Ernst, and Simon Gächter, 2000. "Cooperation and Punishment in Public Goods Experiments", *American Economic Review* 90, 980-994.

Fischbacher, Fong and Fehr (in preparation), "Fairness and Competition", 2000.

Fischbacher, Gächter and Fehr (forthcoming), "Are People Conditionally Cooperative? – Evidence from Public Goods Experiments", *Economic Letters*.

Forsythe, Robert L., Joel Horowitz, N. E. Savin, and Martin Sefton (1994); "Fairness in Simple Bargaining Games," *Games and Economic Behavior* 6, 347-369.

Fudenberg, Drew and Levine, David. The theory of learning in games. Cambridge: MIT Press, 1998

Frey, Bruno and Iris Bohnet. Identification in democratic society. Journal of Socio-Economics, 26, 1997, 25-38.

Friedman, Daniel and Sunder, Shyam (1994); Experimental Methods - A Primer for Economists, Cambridge University Press, Cambridge.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982); "An Experimental Analysis of Ultimatium Bargaining," *Journal of Economic Behavior and Organization* III, 367-88.

Güth, Werner (1995); "On the Construction of Preferred Choices – The Case of Ultimatum Proposals", Discussion Paper Economic Series No. 59, Humboldt University Berlin.

Hoffman, Elisabeth, Kevin McCabe, and Vernon Smith, 1996. "On Expectations and Monetary Stakes in Ultimatum Games," *International Journal of Game Theory* 25, 289-301.

Holm, Hakan J. (2000); "Gender Based Focal Points. Games and Economic Behavior 32 (2), 292-314.

Isaac, Mark R., James M. Walker, Arlington W. Williams (1994) " Group Size and the voluntary Provision of Public Goods", *Journal of Public Economics* 54, 1-36.

Kahneman, Daniel, Jack L. Knetsch, and Richard Thaler (1986); "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review* LXXVI, 728-41.

Ledyard, John (1995) "Public Goods: A Survey of Experimental Research", Chap. 2 in: Alvin Roth and John Kagel (eds.), *Handbook of Experimental Economics*. Princeton: Princeton University Press.

Mehta, Judith, Chris Starmer and Robert Sugden (1994); the Nature of Salience – An Experimental Investigation in Pure Coordination Games", American Economic Review 84, 658-673.

North, Douglass (1990); *Institutions, Institutional Change and Economic Performance*, Cambridge: Cambridge University Press.

Ostrom, Elinor (2000); "Collective Action and the Evolution of Social Norms", *Journal of Economic Perspectives*, 14, 137-158.

Rabin, Matthew (1993); "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83(5), 1281-1302.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir (1991). "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review* 81, 1068-95.

Sally, David (1995); "Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of Experiments from 1958-1992", Rationality and Society 7, 58-92.

Schelling, Thomas (1960); *The Strategy of Conflict*, Cambridge MA: Harvard University Press. 1960.

Smith, Natalie (2000); "Ultimatum and dictator games among the Chaldeans of Detroit". Talk to MacArthur Foundation Anthropology project, December 4, 2000.

Weibull, Jorgen (1995); *Evolutionary Game Theory*, Cambridge: MIT Press.

Williamson, Oliver (1985); *The Economic Institutions of Capitalism,* New York: Free Press.