View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by UCL Discovery

# Experimental Economics: Science or What?

Ken Binmore
Economics Dept
University College
Gower Street
London WC1E 6BT
United Kingdom

Avner Shaked
Economics Dept
Bonn University
Adenauerallee 24
53113 Bonn
Germany

**Abstract:** Do we want experimental economics to evolve into a genuine science? This paper uses the literature on inequity aversion as a case study in warning that we are at risk of losing the respect of other scientific disciplines if we continue to accept the wide claims about human behavior that are currently being advanced without examining either the data from which the claims are supposedly derived or the methodology employed in analyzing the data.

# Experimental Economics: Science or What?

**by Ken Binmore and Avner Shaked**[1]

## 1 Introduction

The long heralded reintegration of economics with psychology is now an accomplished fact. But although experimental economics is now a mainstream activity, it remains an immature discipline that may evolve in various directions depending on how things go in the next few years. Should we join with psychologists in aspiring to a strictly scientific attitude to experimental work? Or should we continue to follow the tradition in policy-orientated economics and treat experiment results as just one more rhetorical tool to be quoted when convenient in seeking to convert others to whatever your own point of view may be?

In this paper, we urge experimentalists to make a conscious break with the less respectable traditions of economic debate and join psychologists in adopting a more skeptical attitude to the far-reaching claims about human behavior that are currently being extrapolated from very slender data. As a case study, we examine some of Fehr and Schmidt's [21, 22, 23, 19, 20] influential papers on the theory of inequity aversion. We find a large discrepancy between the claims that are made on behalf of the theory and the support these claims actually enjoy from the data from which they are supposedly derived.

We appreciate that everyone can name celebrated economists who are vulnerable to criticism on similar grounds, but the issue isn't whether current practice in experimental economics is better or worse than in certain other branches of economics, but whether experimental economics can survive critical comparison with better established scientific disciplines like cognitive psychology. From this point of view, the astonishing popular success that Fehr and Schmidt have enjoyed in recent years inevitably puts them on the front line.

[1]Ken Binmore gratefully acknowledges the financial support of the UK Economic and Social Research Council through the Centre for Economic Learning and Social Evolution at University College London. Avner Shaked gratefully acknowledges the financial support of the Deutsche Forschungsgemeinschaft through SFB/TR 15.

# 2 The Optimizing Paradigm

This section tries to set the record straight on a number of uncontroversial facts that are commonly overlooked in the heat of debate. The opinions we express are, of course, another matter.

**De gustibus non est disputandum.** Behavioral economists sometimes claim that neoclassical economists hold that people are selfish. Henrich *et al* [35] go so far as to assert the existence of a "selfishness axiom".[2] But no such axiom appears in standard economics textbooks. On the contrary, economists are all taught that "there is no accounting for tastes." When utility functions of various kinds are fitted to data obtained in laboratory experiments, neoclassical economics is therefore in no danger of being refuted.

Behavioral economists differ from neoclassical economists on this front only in having seemingly reverted to the classical view that people really do have utility generators in their heads. Neoclassical economists don't deny this possibility, but they follow Paul Samuelson in thinking it a virtue not to be committed to any particular psychological view of how human minds work.

**Maximizing money?** To say that agents are money-maximizers doesn't imply that they are selfish, For example, if Mother Teresa [38] had been a subject in one of Fehr and Schmidt's experiments, she would likely have sought to maximize the money she made with a view to distributing it among the poor and needy. Nor does saying that agents care only for their own well-being imply that they are money-maximizers. We can only identify the utils of neoclassical agents with units of some numeraire when their utility functions are quasilinear. Agents who maximize Cobb-Douglas utility functions are therefore not money-maximizers—nor are they necessarily selfish, since nothing says that they intend to consume the commodity bundles they buy themselves. After all, most people shop on behalf of their households and not just for themselves.

**Maximizing money in games.** Neoclassical economics offers no theoretical support for identifying utils with dollars. Any support for this widespread practice

[2]We quote from Henrich *et al* [35] here and elsewhere, because the book's long list of coauthors includes a representative jury of prominent members of the behavioral school. For a review, see Samuelson [57].

must therefore be empirical. We doubt that many applied workers are aware of the fact, but there is actually a vast amount of experimental evidence from laboratory games that supports the practice in certain circumstances.

All experimental economists accept this claim for market games, but behavioral economists seldom acknowledge that it also holds for most games with money payoffs that have a unique Nash equilibrium —provided that the payoffs are sufficiently large and the subjects have ample time for trial-and-error learning. In spite of much rhetoric to the contrary, the one-shot Prisoners' Dilemma is a case in point. Camerer's [13, p.46] $Behavioral\ Game\ Theory$ says that this fact is too well known[3] for the evidence to require review.

It doesn't follow that there is no room for experienced subjects to exhibit other-regarding preferences in these experiments. It is uncontroversial that most people care about others to some extent. Even Milton Friedman apparently gave money to charity. However, perturbing the utility functions of all the players by introducing a $small$ other-regarding component will not normally move a Nash equilibrium very much. Nor will introducing a $small$ percentage of subjects who care a lot about other people usually affect the data significantly. However, there are games in which theoretical predictions based on money payoffs are not robust to such small perturbations. Public goods games with punishment are an example (Steiner [62]). Theoreticians are at fault when they fail to point out such a lack of robustness, but experimentalists also cannot escape blame if they treat such fragile examples as typical.

We agree with critics of standard practice in empirical economics that the experimental support for modeling agents as maximizers of money extends neither to inexperienced subjects nor to most games with multiple Nash equilibria. Our own view is that inexperienced or inadequately incentified subjects cannot usefully be modeled as optimizers of anything at all. We think they usually begin by operating whatever social norm happens to get triggered by the framing of the laboratory game. As Henrich $et\ al$ [36] say of their (inexperienced) subjects: "Experimental play often reflects patterns of interaction found in everyday life." If this is right, then experimentalists need to look to social sciences other than economics to make sense of the behavior of inexperienced subjects.

**Backward induction?** The case of multiple Nash equilibria is hard because most game theorists regard the equilibrium selection problem as unsolved. How-

[3]Through the surveys of Ledyard [45] and Sally [54].

ever, as in Henrich *et al* [35], it is common in behavioral economics to proceed as though the optimizing paradigm implies backward induction. This mistake matters a great deal, because the games on which behavioral economists currently focus typically have many unacknowledged Nash equilibria. For example, all possible divisions of the money in the Ultimatum Game are Nash equilibrium outcomes. Full cooperation in public goods games with punishment is a Nash equilibrium outcome.

Aumann [1] proves that common knowledge of rationality implies backward induction in finite games of perfect information, but his (controversial) definition of rationality isn't simply maximizing utility. Nor does Aumann [2] recommend using backward induction to predict in laboratories, since he shows that slight perturbations of his conditions can dramatically alter players' behavior. Reinhard Selten–the inventor of subgame-perfect equilibrium—never thought that backward induction would predict in laboratories. It was to demonstrate this fact that he proposed to Werner Güth [33] that he carry out the very first experiment on the Ultimatum Game.

Most game theorists are even more skeptical of backward induction, having abandoned refinement theories more than twenty years ago. Their skepticism was deepened after it was discovered that simple trial-and-error adjustment processes may easily take players in a game to a Nash equilibrium that isn't subgame-perfect, or which is weakly dominated (Samuelson [56]). This is true, for example, of the replicator dynamics in the Ultimatum Game (Binmore *et al* [9]).

Of the very large numbers of experimental papers that refute backward induction in the laboratory, our favorite is Camerer *et al* [13, 15], whose results don't depend at all on what, if anything, the subjects may be trying to optimize. In three-stage Ultimatum Games with alternating offers, subjects commonly don't even click on the screen that would show the final subgame from which a backward induction necessarily begins.

An old experimental result of ours is still sometimes quoted to the contrary, but when the full range of results on two-stage Ultimatum Games is considered, it is clear that backward induction on its own cannot come near explaining the data (Binmore [8]). In a recent paper, we show that even attributing utility functions to the subjects that take account of both player's money payoffs—regardless of the functional form—cannot rescue backward induction in two-stage Ultimatum Games [10]. But like the zombies in horror movies that keep getting up no matter how many bullets are pumped into them, backward induction seemingly cannot be laid to rest.

**Social preferences?** Note finally that to deny that a substantial fraction of people have been shown to have utility functions with a large other-regarding component is not to make any claims at all about what may actually be the case. We certainly don't think that fairness is unimportant in human life. Nor do we think that social preferences are a myth. But inequity aversion is only one of many theories about how and why fairness is mediated through social preferences (Binmore [4, 5, 7]).

# 3 Testing Theories Scientifically

How should a theory of human behavior be tested in the laboratory? We don't think that some new solution to the problem of scientific induction is required. Our unoriginal view is that we should simply follow what is regarded as best practice in other sciences (Guala [31]). One might, for example, seek to emulate the psychological work of Tversky [63]. If this observation seems at all controversial, it is because empirical economists have traditionally faced very different challenges from physicists or chemists. Macroeconomic data is sparse and uncontrolled, and therefore usually consistent with multiple theories. Since its interpretation is often relevant to policy, it is therefore not surprising that we have learned to tolerate advocates of one theory or another talking up their own position and misrepresenting the position of their rivals. It wouldn't be in equilibrium if one party always found itself on the losing side because it never argued beyond its data.

However, the data in experimental economics need neither be sparse nor uncontrolled. Nor are our conclusions often immediately relevant to policy. We can therefore afford to aspire to higher standards than are traditional in mainstream economics, even if we don't always succeed in measuring up to our aspirations.

**1. Prediction.** Experimental papers in economics usually describe the data of an experiment and then propose a theory that fits the data if various parameters are suitably chosen. Sometimes sophisticated econometric techniques are used (although not nearly so sophisticated as Manski [49] recommends).

But the scientific gold standard is prediction. It is perfectly acceptable to propose a theory that fits existing experimental data and then to use the data to calibrate the parameters of the model. But, before using the theory in applied work, the vital next step is to state the proposed domain of application of the

theory, and to make specific predictions that can be tested with data that wasn't used either in formulating the theory or in calibrating its parameters.

For obvious reasons, scientists usually insist that experiments be run anew so that new data can be used to test predictions. We can't do this with macroeconomic field data, but the data gathered in economic laboratories isn't macroeconomic field data. Scientists also attach much importance to replication. One report of a successful prediction is regarded only as provisional until it has been independently replicated in another laboratory. It is literally impossible to replicate a natural experiment in macroeconomics, but laboratory experiments are not natural experiments.

A problem in economics is that editors are reluctant to publish reports of replications. The standard practice in physics of replicating a previous experiment before moving on to a new design is therefore almost absent in economics.

**2. Respecting logic.** A theory usually has many implications that can be tested. If one prediction of the theory is deduced from another, then one cannot claim a success for the theory if the consequent is verified but the antecedent is refuted. One has actually shown that a rival theory must hold that predicts the consequent and the *negation* of the antecedent.

For example, a theory might hold that Chicago economists drink too much. Drunken folk are prone to fall down stairs, but a report that Milton Friedman once fell down some stairs provides no support for the theory if he was known to have been entirely sober at the time.

**3. Cherry picking.** How would one test the hypothesis that Milton Friedman was a bleeding-heart liberal? Surely not by checking that he gave a tiny fraction of his income to charity. The point is that some events are easier to predict than others. A good theory will successfully predict events that are difficult to predict.

Cherry picking the events that a theory is to predict—especially if done after the test experiment has been run—is obviously unacceptable. But how is one to know in advance which events are difficult to predict and which are easy? We think a minimal requirement is to compare the predictions of the theory to be tested with rival theories. If many of the theories predict a particular event, then predicting that event should be deemed to be easy. Cherry picking the rival theories is no more acceptable than cherry picking the events to be predicted.

**4. Predicting or fitting?** Ptolemy's theory of epicycles fits the movement of the planets better than Kepler's ellipses—provided enough epicycles are allowed. It is therefore necessary to be very careful when parameters are left floating and

so are available to be fitted to new data which is supposedly being predicted. Hanson and Heckman [34] is the appropriate authority.

The history of non-expected utility theory provides a good example. Kahneman and Tversky [41] showed that Von Neumann and Morgenstern's theory of expected utility (which has no parameters) is a bad predictor in the laboratory. So various alternative theories were proposed that fitted the data better than expected utility theory when their parameters were suitably chosen. This work generated much enthusiasm, and many applied papers were written incorporating one or another non-expected utility theory. But this literature seems now to have largely dried up after two papers appeared in the same issue of *Econometrica* showing that, when like is compared with like, all extant theories predict badly—but orthodox expected utility theory is the least bad. (See Camerer and Harless [14], Hey and Orme [37]; also a recent paper by Schmidt and Neugebauer [58].)

**5. Honest reporting.** It is important not to remain silent about the successes of rival theories or the failures of one's own theory. If there are floating parameters, their existence should be be frankly acknowledged. The methodology should be clearly explained with a view to assisting replication. Running unreported "pilot" studies is not good practice. Hiding significant information in footnotes or technical appendices is not acceptable.

# 4 Inequity Aversion Theory

The idea of fitting a utility function incorporating inequity aversion to experimental data originates with Bolton and Ockenfels [11, 12]. Player $i$'s utility

$$U_i(x) = U_i(x_1, x_2, \ldots, x_n)$$

in an experimental game is assumed to be a function not only of his or her own money payoff $x_i$ but also of the money payoffs of the other $n-1$ players.

Fehr and Schmidt, 1999 [21] diverge from the theory of Bolton and Ockenfels by postulating a more tractable functional form:

$$U_i(x) = x_i - \frac{\alpha_i}{n-1}\sum_{j\neq i}(x_j - x_i)^+ - \frac{\beta_i}{n-1}\sum_{j\neq i}(x_i - x_j)^+$$

where $0 \leq \beta_i < 1$, $\beta_i \leq \alpha_i$ and $x^+ = \max\{x, 0\}$.

Player $i$ is therefore characterized by a pair of parameters $(\alpha_i, \beta_i)$. The parameter $\alpha_i$ measures player $i$'s envy at being poorer than others. The parameter $\beta_i$ measures player $i$'s discomfort at being richer. Fehr and Schmidt also follow Levine [48] in allowing players to be heterogeneous, so that a population is characterized by a joint distribution of $\alpha$ and $\beta$.

The subject population in the theory of Fehr and Schmidt is heterogeneous, and so the players face some risk, since the parameters of their partners in a laboratory game are uncertain. Fehr and Schmidt deal with this problem by assuming that the players maximize expected utility, as in Bayesian decision theory. The subjective probability distributions of the players are taken to be the actual joint distribution of $\alpha$ and $\beta$, which the players have presumably learned in previous encounters or outside the laboratory. Since the theory assumes that they play (subgame-perfect) equilibria in the laboratory, all this information is taken to be common knowledge.

Range of application? Nobody thinks that laboratory results on the Dictator Game will predict how much people will donate from their wallets to strangers they pass in the street. The range of application of results in the Ultimatum Game is similarly limited. For example, Hoffman $at\ al$ [39] shows that subjects who believe that they have earned the right to propose in the Ultimatum Game expect more. Ball and Eckel [3] show that merely pinning a gold star on some players can have the same effect. Concealing the amount available for division from the responder also increases the amount demanded (Mitzkewitz and Nagel [50]). The anthropological studies reported in Henrich $et\ al$ [35] show that behavior in the Ultimatum Game can differ markedly in different traditional societies.

It is well-established that similar contextual considerations apply more generally to fairness attitudes. Konow [43] is the most recent economist to press this point, but Selten [60] vainly drew our attention to the work of Homans [40] many years ago. It is something of a scandal that we ignore a substantial experimental literature[4] in social psychology, which is said to show that what subjects count as fair depends on a whole range of contextual parameters, including perceived need and prior investment of effort.

A context-free theory like inequity aversion cannot therefore apply to all sit-

[4]The literature offers experimental support for Aristotle's ancient contention that what is fair is what is proportional. For example, Deutsch [16], Kayser $et\ al$ [42], Lerner [46, 47], Reis [52], Sampson [55], Schwartz [59], Wagstaff [64], Walster $et\ al$ [65, 66].

uations in which fairness attitudes matter. We are not very clear on what its proponents regard as the appropriate domain of application, and so we only examine tests proposed by Fehr and Schmidt themselves.

# 5 Fehr and Schmidt's $QJE$ Article

Fehr and Schmidt have written many papers that discuss inequity aversion with and without various coauthors. We can't discuss all this work, and so focus on the foundation stone of the research program, which is a hugely influential paper that appeared in the *Quarterly Journal of Economics* (Fehr and Schmidt, 1999 [21]). However, since the $QJE$ paper refers to the results of several further papers, there remains much ground to cover. Section 9 considers other papers [23, 19, 20] with a view to demonstrating that our methodological complaints would seem to apply to Fehr and Schmidt's whole oeuvre.

It isn't easy to keep track of Fehr and Schmidt's methodology in the $QJE$ paper [21], but they are very clear in a later survey paper prepared for an invited lecture at the World Congress of 2000. In discussing the work that stems from their $QJE$ paper, Fehr and Schmidt, 2003 [22, p.222] say:

> Using the data that is available from many experiments on the ultimatum game, Fehr and Schmidt calibrate the distribution of $\alpha$ and $\beta$ in the population. Keeping this distribution constant, they show that their model yields quantitatively accurate predictions across many bargaining, market and co-operation games.

Fehr and Schmidt would therefore seem to endorse the predictive criterion outlined in the previous section. However, we shall argue that they actually make good on none of their claims in this passage. In particular:

**1.** The Ultimatum Game data that is supposedly used to calibrate the parameters $\alpha$ and $\beta$ is inadequate for this purpose. It is logically impossible to tie down the parameters from the data that is supposedly used to estimate them. We are then left with floating parameters which could later be fitted to new data that their model is said to predict. For this reason, we shall speak of a parametrization rather than a calibration when referring to the manner in which Fehr and Schmidt assign values to their parameters in the various papers we examine.

**2.** Fehr and Schmidt don't keep the distribution of parameters constant. This is a major complaint, since fitting a model to new data is not at all the same thing

as predicting new data with a model whose parameters have been calibrated with existing data.

**3.** Fehr and Schmidt don't obtain "quantitatively accurate predictions" across many games. In the $QJE$ paper, they analyze four games. In only one game does their model arguably predict significantly better than a money-maximizing model, and even here the case is unclear because they cherry pick both what they choose to regard as a prediction, and the variant of a money-maximizing model with which their model is compared.

**4.** In a reply to an earlier critique on Shaked's website, Fehr and Schmidt [26] later say that it was their results on contract games that gave them the confidence to make such large claims for their calibrated $QJE$ model.[5] We consider these contract games in Section 9, but conclude that the data refutes their claims.

# 6 Calibrating the Distribution of Parameters?

Fehr and Schmidt, 2003 [22, p.222] claim to have used data from Ultimatum Game experiments to calibrate their inequity-aversion model in their $QJE$ paper. The marginal and joint distributions they claim to use in their parametrized model are given in Table 1.

How is Table 1 calculated? According to Fehr and Schmidt's theory of inequity aversion, the behavior of inequity-averse responders in the Ultimatum Game is solely determined by their envy parameter $\alpha$, and that of proposers by their discomfort parameter $\beta$. So the marginal distribution of $\beta$ needs to be calculated from data on the proposers' offers, while the marginal distribution of $\alpha$ needs to be calculated from data on the responders' acceptance rates.

However, if the inequity aversion theory holds in the Ultimatum Game, all that can be said about proposers who made an offer of 50% or more of the available money is that their $\beta$ exceeds $0.5$ [21, Proposition 1, p.826]. But it will be noted that 40% of subjects have been assigned a value of $\beta = 0.6$ in Table 1.

The Ultimatum Game data that would be needed to estimate the joint distribution of $(\alpha, \beta)$ seems to be absent altogether. For this purpose, we would

[5] Their reply also cites two further papers that aren't mentioned in their invited address [22]. We don't discuss Fischbauer, Fong and Fehr [28] because it depends on quantal response equilibria, whose explanatory value remains controversial. Fehr and Schmidt [24] doesn't seem relevant to their $QJE$ parametrization at all.

| $\alpha$ | % |
|---|---|
| 0.0 | 30 |
| 0.5 | 30 |
| 1.0 | 30 |
| 4.00 | 10 |

marginal distribution

| $\beta$ | % |
|---|---|
| 0.0 | 30 |
| 0.25 | 30 |
| 0.6 | 40 |

marginal distribution

| $(\alpha, \beta)$ | % |
|---|---|
| (0.0,0.0) | 30 |
| (0.5,0.25) | 30 |
| (1.0,0.6) | 30 |
| (4.0,0.6) | 10 |

joint distribution

Table 1: Distributions of $\alpha$ and $\beta$ in the population of subjects taken from the $QJE$ paper—supposedly calibrated from Ultimatum Game data. The marginal distributions appear in their Table III. The vital joint distribution is to be found only at the very end of their appendix, where Fehr and Schmidt say that they are assuming perfect correlation "for concreteness" although their assumptions are "clearly not fully realistic" [21, p.846].

need information on how each individual subject behaved $both$ as a responder $and$ as a proposer. We were unable to locate such information in the quoted source papers on the Ultimatum Game. When Fehr and Schmidt say that there is empirical support for assuming correlation between $\alpha$ and $\beta$, they therefore presumably have another source in mind [21, p.864].

However, there seems no point in pursuing this point, since Fehr and Schmidt no longer claim to have calibrated their parameters from Ultimatum Game data. In the $Handbook\ on\ Reciprocity,\ Gift\text{-}Giving\ and\ Altruism$, they have reverted to the language of their $QJE$ article, and now only claim to have chosen their parameters to be $consistent$ with the Ultimatum Game data [27, p.26].

In any case, it is clear that Fehr and Schmidt didn't estimate their parameters from Ultimatum Game data alone. The parameters are under-identified by this data and so can float to a considerable extent. To avoid suspicion of having allowed their remaining freedom of choice to have been influenced by the data that was supposedly to be used to test their parametrized model, it would have been wise of Fehr and Schmidt to have explained their choice methodology in advance. Perhaps some econometrics would have been in order.

As things stand, they continue to claim that their choices of the floating

parameters were made without reference to the data they were planning to predict. For example, in a reply to criticism on Shaked's website, Fehr and Schmidt, 2005 [26, p.7] comment on their decision to adopt a model in which 40% of the population have $\beta = 0.6$ by saying:

> Thus, the condition of Proposition 5 requires $\beta_i \geq 0.6$. We had picked the highest possible value of $\beta_i$ to be $\beta_i = 0.6$ in Table III, which is just sufficient, but very tight.

If they had chosen the highest value of $\beta$ in their model to be 0.55, their model wouldn't have been consistent with the data from the Public Goods Game with Punishment [17] of Section 8.2. If they had chosen the highest value of $\beta$ to be 0.85, their model wouldn't have been consistent with the Competition among Responders Game [32] of Section 8.3.

# 7 Distribution of Parameters Kept Constant?

Of the four games analyzed in the $QJE$ paper, the most striking is the Public Goods Game with Punishment—the game for which it is essential that $\beta \geq 0.6$ (Fehr and Gächter [17]). Fehr and Schmidt's analysis of this game is based on Proposition 5, the proof of which appears at the very end of their appendix. It is only then that we learn of the assumptions that they are making that lead to the joint distribution of $\alpha$ and $\beta$ in our Table 1. However, the proof of Proposition 5 would seem to assume that all the types who aren't 'conditionally cooperative enforcers' (with $\beta \geq 0.6$ and a correspondingly high $\alpha$) are money-maximizers (with $(\alpha, \beta) = (0, 0)$). Perhaps this deviation from their official distribution of parameters is only to simplify the proof, but it signals that Fehr and Schmidt feel free to alter their assumptions about the joint distribution of $\alpha$ and $\beta$ without drawing attention to this fact.

More seriously, Fehr and Schmidt also claim to use the $QJE$ parametrization when making predictions in the three contract papers reviewed in Section 9. However, we shall see that they use the $QJE$ parametrization in none of these papers. Fehr and Schmidt are therefore not only engaged in a *fitting* exercise rather than a *predicting* exercise, they also feel free to change the parameters of their model when new data needs to be accommodated.

# 8 Quantitatively Accurate Predictions?

In this section we consider the four games analyzed in the $QJE$ article with a view to assessing Fehr and Schmidt's claim that their model generates quantitatively accurate predictions.

## 8.1 Public Goods Games without Punishment

ection 2 mentions Camerer's [13] endorsement of the conclusions reached by Ledyard [45] and Sally [54] from their surveys of very large number of experimental studies on Public Goods Games (without punishment). In such a game, the subjects privately choose how much to contribute to a public project that enhances the value of the total contribution. The benefits of this enhanced value are enjoyed by everyone, including the free riders who contributed nothing. The one-shot Prisoners' Dilemma is the best-known example. About 51% of inexperienced subjects cooperate in the one-shot Prisoners' Dilemma, but only about 10% are still cooperating after ten trials or so (Ledyard [45, p.172]).

It is therefore surprising that Fehr and Schmidt,1999 [21, p.818] quote Ledyard's survey early in their $QJE$ paper as though it were hostile to the money-maximizing hypothesis. A good case could be made for using Ledyard's data for this purpose in the case of inexperienced subjects, but we shall find that Fehr and Schmidt test their own theory on experienced subjects. They therefore cannot hope to fit the data much better than the money-maximizing model even though they have a potentially infinite number of extra parameters with which to play, because it is already established that the money-maximizing model fits this kind of the data rather well.

However, rather than quote the conclusions of Ledyard's comprehensive survey of papers published before 1995, Fehr and Schmidt, 1999 [21, p.838] choose for themselves what experimental studies on Public Goods Games to report in their Table II. The particular feature of their choice to which we wish to draw attention is that the table indiscriminately mixes both papers in which the subjects played repeatedly against strangers (the stranger design) and papers in which they played repeatedly with the same partners (the partner design), as though this feature of an experimental design were irrelevant [21, Footnote 18]. However, it is well-known that laboratory subjects commonly play more cooperatively

in the partner design,[6] even when it is common knowledge that only a fixed number of games are to be played.

The leading authority here is Reinhard Selten [61]. It is true that if backward induction hadn't been overwhelmingly refuted in the laboratory, one could treat the final game as though it were a one-shot game, but backward induction has been overwhelmingly refuted—not least by Selten who invented the idea. If one wants to study the behavior of experienced subjects in a one-shot game, the standard experimental technique is the stranger design.

None of the preceding observations are controversial. The facts are even apparent in Fehr and Gächter [17], which is the key source of data for Fehr and Schmidt's $QJE$ paper. No econometrics is necessary to check that one gets different results from the partner design than the stranger design; nor that inexperienced subjects don't always play in the same way as experienced subjects. Just eyeball the figures in Fehr and Gächter [17] (particularly Figure 2 on page 987 and Figure 4 on page 980).

In their $QJE$ paper, Fehr and Schmidt could have chosen to predict at least four pieces of data from Fehr and Gächter [17], depending on whether they chose the stranger or partner design, and whether they chose experienced or inexperienced subjects.[7] They would have got most cooperation in the game $without$ punishment by choosing the partner design with inexperienced subjects. However, they would then have gotten less cooperation in the game $with$ punishment.

In the game with punishment, we shall see that they cherry pick what they choose to predict by taking the case of partner design with experienced subjects. If they had chosen to predict the same case in the game without punishment, their theory would need to have explained the existence of about 51% experienced subjects who free-ride by contributing nothing.[8] Since their parametrized model predicts that nearly 100% of subjects will free ride (Proposition 4 of the $QJE$ paper), the discrepancy is very large.

However, Fehr and Schmidt choose not to predict behavior in the game without punishment for the same case as in the game with punishment. They

[6]Fehr and Schmidt [21, Table II] confuse the issue by saying that all the papers they list are repeated games, but it is only orthodox to speak of a repeated game in the case of play against the same partners.

[7]We neglect the possibility of taking into account whether the game without punishment is played before or after the game with punishment, although the data shows clear differences.

[8]The percentage of 51% is estimated from Figure II of the $QJE$ article, which reproduces Figure 4 in [17]. A precise figure does not seem to be given.

don't even predict behavior in the game without punishment for the case of experienced subjects in the stranger design (when 51% is replaced by 75%). They choose to predict data that doesn't appear in the paper of Fehr and Gächter [17] at all. They compare the prediction of their theory (nearly100% free riders[9] ) with the average percentage of 73% obtained from their Table II (which summarizes data from a selection of other experiments). One would have thought that even this gap between their prediction and the data supposedly being predicted would give them pause, but Fehr and Schmidt, 1999 [21, p.845] say:

> Thus, it seems fair to say that our model is consistent with the bulk of individual choices in this game.

A footnote is appended that seeks to explain away the difference of 27% (not 25% as they say) between the number they choose to report and their prediction, but it was Fehr and Schmidt who chose to make the percentage of free riders the basic criterion.

To summarize, Fehr and Schmidt, 1999 [21] potentially misrepresent the extent to which a money-maximizing model can explain the data in Public Goods Games without Punishment by selectively quoting from the literature. They then do not predict the data of Figure II (Fehr and Gächter [17, Figure 4]) as would be appropriate given that this is what they do for games with punishment. Instead they predict data from their selection from the literature. They then claim success, even though there is a gap of 27% between their prediction and the criterion they choose to examine. If they had predicted the data for experienced subjects in the partner-design case from Fehr and Gächter (Figure II in the $QJE$ paper) as they do for games with punishment, the gap would have been about 49%. Finally, it is tendentious to be considering the partner design at all, when it is standard practice to use the stranger design—especially when they don't make it clear that they are deviating from standard practice.

There is a certain irony in all this twisting and turning, because the predictions of the money-maximizing model and Fehr and Schmidt's parametrized model are nearly the same (100% free riders). If compared using the same data, they will therefore perform nearly as well as each other.

[9] Fehr and Schmidt's actual prediction of the percentage of free-riders varies with the experiment considered. The average across all experiments (weighted by sample size) is 98.48%.

## 8.2 Public Goods Game with Punishment

This game is the only case analyzed in the $QJE$ paper in which the question isn't whether inequity aversion can predict as well as (or better than) the money-maximizing model in cases when the latter is acknowledged to predict rather well for experienced subjects.

In the first stage of the Public Goods Game with Punishment, each subject can simultaneously contribute towards a common pool, whose value is then enhanced and eventually redistributed to all players (including any free riders). This standard Public Goods Game is then modified so that the subjects can punish each other at a second stage. Each player is informed of the contributions of the others, and then has the opportunity to reduce the payoff of a selected victim by 10% on payment of a small cost. A money-maximizing player in the one-shot game will never punish because there is nothing to gain by changing the behavior of a person who will never be encountered again. However, Fehr and Gächter [17] confirm Yamagishi's [67] finding that free riders do get punished, and that the average level of contribution is relatively high for experienced subjects.

We agree that this is a striking result that requires explanation. The question is how well Fehr and Schmidt's parametrized model does in predicting the result as compared with other theories.

We first review Fehr and Schmidt's methodology when arguing in favor of their parametrized model of inequity aversion. We have already mentioned that a value of $\beta$ of at least 0.6 needs to be attributed to subjects who offer 50% of the available money in the Ultimatum Game in order to find a parametrization of the inequity aversion model that is consistent with the data from the Public Goods Game with Punishment—although no evidence is available to support this choice. One must also go beyond the Ultimatum Game data in postulating a substantial correlation between $\alpha$ and $\beta$.

We have also drawn attention to the fact that Fehr and Schmidt cherry pick which set of data they will predict. Standard practice[10] would call on them to predict the data from the stranger design given in Figure 2 of Fehr and Gächter [17] but they choose instead to predict the data from the partner design given in Figure 4 (reproduced as Figure II in the $QJE$ paper). The two sets of data are very different. For example, somewhat more than 80% of experienced subjects contribute the maximum in the partner design, but only slightly more than 10%

[10] For example, the three contract papers considered in Section 9 use the stranger design.

contribute the maximum in the stranger design. It isn't clear why they indulge in this cherry-picking exercise, since our impression is that the data from the stranger design would still present a major challenge to what they regard as the unique money-maximizing prediction (in which all subjects free ride).

But are they right to proceed as though the $unique$ money-maximizing prediction is that all subjects will free ride? In the case of the partner design—when the same four subjects knowingly play each other ten times, with only the final round being predicted—the answer is certainly $no.$

There are two reasons. The first is that the subjects have the opportunity to learn the types of their partners during the nine repetitions of the game that precede the final game whose data is predicted. For example, there is a positive probability that all four partners in a session will turn out to be money-maximizers. It might be argued that the subjects are too naive to update their initial beliefs about the types of their partners, but there is a more important reason for not using the partner design.

In pursuing this second point, first note that Fehr and Gächter's [17] results for their fixed-horizon game are consistent with those of Ostrom $et\ al$ [51] for the case of an indefinite horizon (in which case the folk theorem of repeated game theory says that almost any outcome can be supported as a subgame-perfect equilibrium). The well-known Gang of Four paper (Kreps $et\ al$ [44]) explains why attributing irrational behavior to just a small fraction of the population can generate the same conclusions in a game with a finite horizon as in the case with an indefinite horizon. It is therefore not surprising that Selten [61] is only one of many authors who have found that experimental behavior in finite-horizon games is often close to the behavior predicted for the corresponding infinite-horizon game. In the case of the Public Goods Game with Punishment, Steiner [62] has even written a model in which a slight perturbation to the money-maximizing paradigm is enough to generate a subgame-perfect equilibrium in the ten-times repeated version with four players in which everybody contributes the maximum.

Cherry picking the data from the partner design is therefore only an effective tactic for Fehr and Schmidt if they are also allowed to cherry pick the rival money-maximizing prediction to be one that denies that a $small$ fraction of the subject population will do something other than maximize money. But not even the most determined Chicago redneck would want to deny the existence of some small fraction of deviants from the money-maximizing paradigm.

But what if Fehr and Schmidt had not cherry picked the partner design, but followed standard practice in predicting the stranger design? They would

still not be entitled to identify the money-maximizing prediction with the unique subgame-perfect equilibrium in which everybody acts as a free rider. The money-maximizing paradigm arguably entails the play of a Nash equilibrium by experienced players, but nothing says that even experienced players will have learned to play Nash equilibria in subgames that are unreached in equilibrium (as required by standard theoretical arguments offered in defence of subgame-perfection). At the same time, the laboratory evidence is implacably hostile to backward induction, whatever preferences are attributed to the subjects. So what Nash equilibria are available in the one-shot Public Goods Game with Punishment? The answer is that any pattern of contributions can be supported as Nash equilibria in the game with money-maximizing players—including the case when all players contribute the maximum. The players plan to punish any deviation and the fact that such punishment is irrational is undiscovered because nobody deviates for fear of being punished. There will, of course, be deviations by inexperienced players who will be punished by other inexperienced players, but there is no particular reason why the learning process should settle on a subgame-perfect equilibrium rather than one of the many alternative Nash equilibria.

Not only does the money-maximizing model have multiple equilibria, the same is true of Fehr and Schmidt's inequity-aversion model. According to their own analysis, their model admits a continuum of equilibria. It turns out that any level of contribution whatever can be defended as the outcome one of these equilibria. However, Fehr and Schmidt, 1999 [21, p.842] cherry pick the equilibrium they will use for predictive purposes, saying

> Hence, this equilibrium is a natural focal point that serves as a coordination device even if the subjects choose their strategies independently.

In summary, Fehr and Schmidt tell us that they predetermined their floating parameters in a manner that turned out to fit the Public Goods Game with Punishment. They cherry pick the data they choose to predict, the equilibrium from their own model they choose to treat as their prediction, and the equilibrium from a money-maximizing model which they choose to treat as the rival prediction. Insofar as they advocate modeling subjects as players with other-regarding utility functions who honor the backward induction principle, they fail to mention either the theoretical objections to this proposal or the laboratory evidence that militates against it.

## 8.3 Two Auctioning Games

It is uncontroversial that the money-maximizing paradigm works well in predicting the play of experienced subjects in market games. It is therefore not surprising that the money-maximizing paradigm also works well in auctioning games that are not too complicated, since a market can be viewed as a institution in which both buyers and sellers participate in a (formal or informal) auctioning process. Fehr and Schmidt's purpose in considering two auctioning games in their $QJE$ paper was presumably to demonstrate that their model of inequity aversion performs no worse than a money-maximizing model. This is not a very high hurdle to jump. After all, it is famous that even modeling the subjects in laboratory markets as 'zero-intelligence' traders[11] is quite successful in predicting final outcomes (Gode and Sunder [30]).

**Market with Competition among Proposers.** The predicted data comes from a paper by Roth $et\ al$ [53]. A number of proposers make offers to a single responder who must accept or reject the highest offer. One of the proposers who made the highest offer is randomly chosen to divide the surplus with the responder. Sufficiently experienced subjects end up implementing the competitive outcome, in which the proposers offer all the surplus to the responder.

This is the only experiment that we have examined in which the Fehr-Schmidt theory of inequity aversion fully explains the data, perhaps because the prediction is independent of the parametrization. Of course, the money-maximizing model also explains the data equally well. Many other models would also suffice for this purpose. If the responder had been allowed the freedom to choose an equitable offer (rather than being forced to consider only the highest offer), the experiment would have provided a test of Fehr and Schmidt's parametrized model, but all that can be said with the current data is that their model predicts as well as any other alternative.

**Market with Competition among Responders.** The predicted data comes from a paper of Güth $et\ al$ [32]. A proposer makes a single offer to a number of responders. A responder is then selected at random from among those who accepted the offer to divide the surplus with the proposer. The acceptance threshold of responders quickly converged to the very low levels predicted

[11] Who honor their budget constraint, but otherwise bid at random.

by an orthodox competitive analysis.

Fehr and Schmidt's Proposition 3 shows that their model has a continuum of subgame-perfect equilibria of which one generates the same prediction as the money-maximizing model provided that $\beta < \frac{5}{6}$. Since the maximum value of $\beta$ in their parametrization is 0.6, this requirement is easily accommodated. However, nothing in the Ultimatum Game data restricts the maximum value of $\beta$.

In summary, Fehr and Schmidt's inequity aversion model is no worse as a predictor of the two auctioning games than the money-maximizing model. It is also no worse than the money-maximizing model in the case of the Public Goods Game without Punishment. But these are not decisive tests, because the same could be said of a great variety of models. The only real test in their $QJE$ paper is provided by the Public Goods Game with Punishment, where we take issue with more or less everything they say.

# 9 Three Contract Games

In a reply to Shaked, Fehr and Schmidt [26] say that it was the results of the three contract games examined in this section that were decisive in their deciding to make the the far-reaching claims quoted in Section 5. However, we find no support for this claim.

The three contract papers have the common feature that a principal offers a contract to an agent. The agent can then exert a costly effort, which generates a payoff for the principal. Section 8.2 criticizes Fehr and Schmidt's decision to focus on the partner design in the Public Goods Game with Punishment, but all three contract papers use a stranger design. More experiments are reported, but we concentrate on those in which players choose between:

**1.** Bonus, trust, and incentive contracts (Fehr, Klein and Schmidt [19]). In a bonus contract, the principal names a wage, an effort level and a bonus, which she may or may not later pay. A trust contract is the same, except that the final stage in which a bonus may be offered is absent. Neither the agent's effort nor the principal's bonus are contractually enforceable. In an incentive contract, the principal may invest in a verification technology. With this in place, she names a wage, demands an effort level, and specifies a fine if the agent's effort falls below this level. The terms and language of this paper also apply to the other contract papers, with only slight variations.

**2.** Joint ownership contracts (equivalent to bonus contracts) and contracts

in which one player intially owns the whole project but can transfer half the ownership rights to the other player (Fehr, Kremhelmer and Schmidt [20]).

**3.** Piecewise and bonus contracts (Fehr and Schmidt, 2004 [23]).

The subjects' behavior over bonus contracts obviously provides the most suitable data for testing Fehr and Schmidt's parametrized theory. Their behavior with other contracts is restricted in a manner that prevents their giving full expression to any inequity aversion built into their utility functions.

## 9.1 Keeping the Distribution Constant?

Fehr and Schmidt don't keep the $QJE$ parametrization constant as they claim. In all three contract papers, the population is instead assumed to consist of 60% money-maximizing individuals with $(\alpha, \beta) = (0, 0)$ (as in the proof of Proposition 5 of the $QJE$ paper) and 40% inequity-averse individuals who all have the same $\alpha$ and $\beta$ that both exceed 0.5. We refer to such a new distribution as a 40-60 distribution. Such a distribution isn't consistent with Ultimatum Game data because Fehr and Schmidt eliminate 30% of the types in Table 1—those types with $\alpha = 0.5$ and $\beta = 0.25$.

Fehr and Schmidt variously justify their use of a 40-60 distribution by saying that they are "following" the $QJE$ calibration, or that the 40-60 distribution is "in accordance" with the $QJE$ calibration, or that the 40-60 distribution is "a simplification" of the $QJE$ calibration ([23, p.470], [20, p.22], [19, p.144]). But it would be a mistake to take these observations to mean that lumping the types eliminated from Table 1 in with the money-maximizers is acceptable on the grounds that they behave no differently from money-maximizers in the three contract games. In the equilibria proposed as explanations of the data, agents (not principals) of the types eliminated from the $QJE$ distribution would $not$ behave like the money-maximizers with whom they have been included.

In Fehr, Kremhelmer and Schmidt [20], and in Fehr, Klein and Schmidt [19] all the higher values of $\alpha$ are equated to 2, although the value $\alpha = 2$ appears nowhere in our Table 1. Other deviations from the $QJE$ parametrization appear in an appendix [25] to Fehr and Schmidt, 2004 [23]. Readers of the published paper will be unaware of these deviations, since the appendix is unpublished.

## 9.2 Quantitatively Accurate Predictions?

In contrast with Fehr and Schmidt's later claims, the contract papers themselves nearly always say that the experiments confirm the 'qualitative predictions' of the model. But since they have been cherry picked, even the limited support they offer for the theory is doubtful.

An exception leads the authors to observe repeatedly that their theory provides "surprisingly accurate quantitative predictions" (Fehr Klein and Schmidt, 2007 [19, pp.123,151]). This claim refers to the fact that the average wage offered in a bonus contract, the average bonus, and the average effort level are close to the averages predicted by the inequity aversion model with a 40-60 distribution. The authors do not conceal that the underlying distribution of the data from which the averages are computed fails to come anywhere near the predictions of the model.[12]

For example, the wage paid and the effort depend on the fractions of principals and agents who play fair, and these fractions differ markedly from their predicted values. We therefore have a case in which the antecedent of an implication within their theory is refuted but the consequent is verified. However, Fehr, Schmidt (and Klein) [19] unconsciously echo Milton Friedman's [29] defense of the Chicago ethos by observing that the subjects behave "as if" motivated by inequity aversion, and that their theory "helps to organize and interpret the data."

We have checked out the extent to which the data relating to bonus contracts is consistent with predictions of the 40-60 model that haven't been cherry picked. This isn't a difficult activity, since there are only four possible types of encounter between a principal and an agent when subjects can only be of two types: "money-maximizing" or "inequity-averse". For example, Table V of Fehr, Klein and Schmidt 2007 [19, p.140] shows that agents whose effort level was at least 5 received an average bonus of of 14.26.[13] But their model predicts a bonus of $0.4 \times 25 = 10$ with a 40-60 distribution. The actual value therefore exceeds the predicted value by more than 42%.

Figure 1 sketches the conclusions of our comparison of the data with the theory, which we document in an appendix. The firmly drawn arrows indicate logical

[12] Recall that Camerer [13, p.46] warns against using average behavior as a summary statistic in Public Goods Games, since subjects tend to split into those who contribute a lot and those who contribute nothing.

[13] Another table, with pooled data of further treatments, presented in the unpublished appendix [18, p.13] yields an average bonus of 13.

implications. Since the predictions of the theory that are verified are implications of more primitive propositions of the theory that the data refutes, one cannot count them in support of the theory. To claim a success for the parametrized theory on the basis of these results is cherry picking. The experiments actually point to the need to formulate some new theory. This might perhaps be a reparametrized version of Fehr and Schmidt's inequity-aversion model. However, our appendix documents why we think this is an unpromising approach.

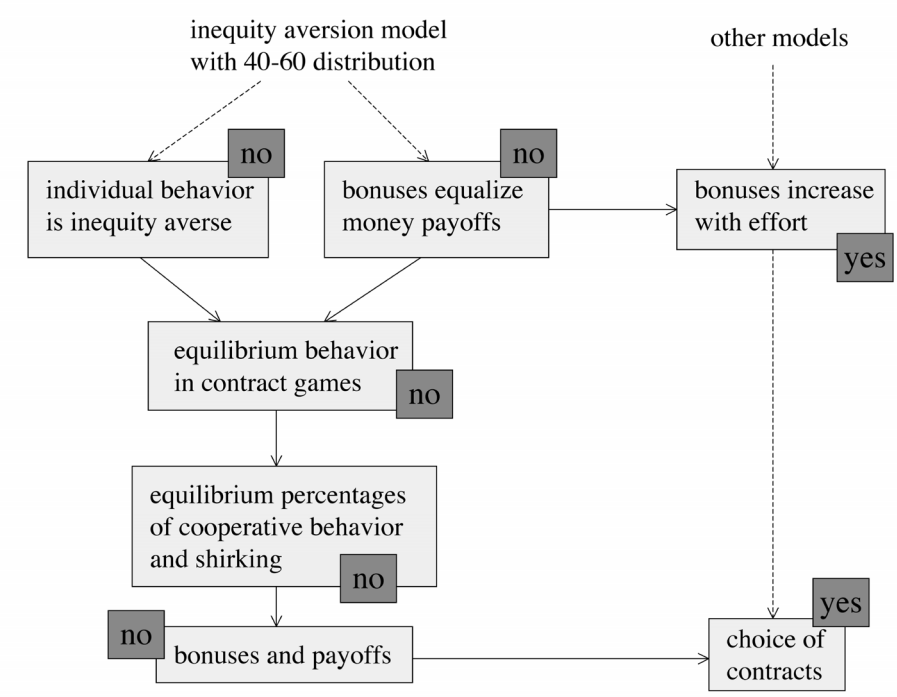


Figure 1: Cherry picking in the three contract games. The boxes marked NO indicate predictions of the inequity aversion model with the 40-60 distribution that the data refutes. The boxes marked YES indicate predictions that are said to be qualitatively accurate. These predictions would follow from a variety of models. The firmly drawn arrows indicate the logical structure of the predictions.

The alternative approach we favor explains the behavior of subjects in terms of social norms. We think it likely that people enter laboratories primed with a variety of social norms, one of which is triggered by the manner in which the experiment is framed. If the resulting behavior is close to a Nash equilibrium of the game (as in the Ultimatum Game), then the social norm is stabilized in

the laboratory environment. If it is not (as in the Prisoners' Dilemma), then the subjects' behavior moves towards a Nash equilibrium. (See Binmore's [6] review of Henrich $et\ al$ [36].)

One would have thought that the progress of the research project of Henrich $et\ al$ [36] (of which Fehr is a part) would have redirected attention at this unremarkable possibility. The anthopological studies in their book uncontroversially debunk the idea that we are genetically programmed with a universal other-regarding utility function. The authors have therefore redirected their attention to confirming that the culturally determined behavior observed in the anthropological experiments is correlated with the extent to which the cultures involved operate market economies.

In summary, Fehr and Schmidt, 2003 [22, p.222] are not entitled to claim that their "calibrated" model yields "quantitatively accurate predictions" in the three contract games, even if they are allowed to alter the $QJE$ parametrization to a 40-60 distribution. On the contrary, their data would seem to $refute$ their parametrized model. A rival model must therefore be sought. We think an obvious alternative is an explanation in terms of social norms.

# 10 Conclusion

Since our motives will be impugned, it is necessary to say that game theorists like ourselves have nothing to fear from any research which genuinely shows that many people can usefully be modeled as having a personal utility function with a large other-regarding component. Our methodology remains unchanged whether our players are Attila the Hun or St Francis of Assissi. We simply recognize that they have different tastes by writing different numbers in their payoff matrices.[14] We are frustrated by the fact that some behavioral economists ignore all denials when claiming that game theory predicts that backward induction will be observed in the laboratory, but not enough to write a paper like this.

What we do care about is the future of our profession. We thought that experimental economics was developing into a science, and we are deeply distressed to find that our aspirations are not as widely shared as we believed.

[14] Even Binmore's [7] theory of fairness assumes nothing about personal preferences.

# A Appendix: Testing Inequity Aversion

This appendix offers some predictions of Fehr and Schmidt's inequity aversion theory with a 40-60 distribution that are refuted by the data. We call the two types that appear in such a distribution money-maximizing or inequity-averse.

Our tests assume that the empirical behavior follows the suggested equilibrium, so that the percentages of inequity-averse principals and inequity-averse agents in the population can be estimated. The fraction of agents who expend low effort is an estimate for the fraction of inequity-averse agents. The fraction of inequity-averse principals can be estimated by computing the fraction of principals who rewarded agents who exerted high effort. We then check whether the estimated distribution agrees with the parametrized model. When we compare the theory's basic predictions with the data we find large disagreements between the two. The estimated fractions of inequity-averse agents and inequity-averse principals are inconsistent with anything close to a 40-60 parametrization.

In all of the contract experiments the data suggests that the number of inequity-averse principals (who pay the bonus) is much larger than the number of inequity-averse agents (who shirk), the ratio between them varies significantly between the three experiments. In Fehr, Klein and Schmidt, 2007 [19] the number of inequity-averse principals needs to be 1.6 times that of the inequity-averse agents, in Fehr and Schmidt, 1994 [23] it needs to be 3 times as large, and in Fehr, Kremhelmer and Schmidt 1995 [20] it needs to be 13 times as large. Any parametrizations that reconcile the data with the theory will therefore need to differ substantially between the experiments. One might perhaps relax the maintained perfect correlation between $\alpha$ and $\beta$ (by introducing individuals with high $\beta$ but low $\alpha$), or postulate a more broadly based distribution of $\alpha$ among the inequity-averse individuals. In either case, the new parametrization will no longer resemble the 40-60 distribution.

Other basic predictions also fail. For example, the bonuses paid in Fehr and Schmidt, 1994 [23] and Fehr, Klein and Schmidt, 2007 [19] don't equalize the payoffs of the principal and the agent.[15] In Fehr and Schmidt, 1994 [23], the

[15] In Fehr, Kremhelmer and Schmidt [20] about 80% of the $A$ players (the principals) equated the payoffs when player $B$ (the agent) made a high effort. Perhaps this is because, in this game, player $A$ receives a high payoff anyway (more than 88), and compensating player $B$ is rather cheap. When player $A$ pays one unit to compensate player $B$, the latter receives 11 units. In the other contract papers, the transfer between the players is one-to-one; each unit player $A$ pays adds one unit to player $B$.

empirical average payoff of an agent is double the predicted payoff, and at least 25% of the agents don't act according to the parametrized theory.

## A.1 Fehr and Schmidt, 2004 [23]

According to the equilibrium for the bonus contract in Fehr and Schmidt, 1994 [23], all principals pay a wage of $w = 225$, money-maximizing agents expend a high (total) effort of $e = 20$, and are paid a bonus of $350$ by the inequity-averse principals. Inequity-averse agents expend a (total) effort of $e = 12$, and are paid no bonus. Proposition 2 of the unpublished appendix [25, p.A-10] lists all pooling equilibria for any permissible fractions $q$ of inequity-averse players in the population, Proposition 3, selects one of these with the help of a refinement-like argument (Condition 1 in p. A-11).

The data is incompatible with the theory in many ways. According to the theory, an inequity-averse agent should make a low total effort of $e = 12$. We find that the fraction of those who made an effort satisfying $10 \leq e \leq 14$ is $55/261 = 21\%$, (Table 2, p. 463), which would imply that the percentage of inequity-averse agents is also 21%.

There is no detailed data on the paid bonuses in the paper, we therefore use the following method to estimate the percentage of inequity-averse principals in the population. The average bonus paid for high effort levels of $18$ or more is about $211$ (Table 2 on p.463 and Figure 3 on p.464):

$$\frac{24}{24+6+63} \times 120 + \frac{6}{24+6+63} \times 170 + \frac{63}{24+6+63} \times 250 = 211.29\,.$$

The estimated fraction of inequity-averse principals in the data is therefore $211/350 = 60.3\%$. The estimated number of inequity-averse agents is 21%, suggesting that there are about three times as many inequity-averse principals as inequity-averse agents, which is incompatible with the 40-60 parametrization.

We also compare the empirical average payoff of an agent with the prediction of the theory. The equilibrium pre-bonus payoff of a money-maximizing agent is 75, and the payoff of a inequity-averse agent is 155. The theoretical average payoff of an agent is therefore $(75+350q)(1-q)+155q$, where $q$ is the fraction of inequity-averse players. The maximum value of this function is 207 (at $q = 0.61$)). It follows that the maximum possible theoretical average payoff of an agent is about half the empirical value, which is about 400 (Figure 4 on p.467).

We also test whether bonuses are paid to equalize the payoffs of the principal and the agent as predicted by the theory. Let $v$ be the production function and $c$ be the cost functions for effort. Let $w$ be the wage and $b$ the bonus paid. Equating the payoffs of the principal and agent implies that $b = (v + c)/2 - w$. Given the frequencies of pairs of efforts in the experiment (Table 2, p.463), we can calculate the average $(v+c)/2$ for any range of efforts. There is no information in the paper about the wage paid, and so we favor the theory by taking it to be the theoretical equilibrium wage, which is $w = 225$. For the range of efforts in which the principal can pay a bonus when $w = 225$,[16] the average bonus that would be needed to equate the payoffs of the agent and principal is 245.77. The average bonus actually paid for the same range of efforts is 180.73 (Table 2 on p.463 and Figure 3). The discrepancy is 36% ($(245.77 - 180.73)/180.73 = 0.3598$)). Even if we require the agent's payoff to be only 80% of the principal's, the discrepancy remains about 15%. Clearly the bonuses were not paid in order to equalize payoffs.

## A.2 Fehr, Klein and Schmidt, 2007 [19]

With the parametrizaion considered by the authors, the inequity-aversion model predicts that all principals will pay a wage of $w = 15$. Money-maximizing agents will make an effort of 7, and be paid a bonus of $25$ by inequity-averse principals. Inequity-averse agents will make a low effort of $e = 3$ and receive a bonus of $1$ from the inequity-averse principals. Money-maximizing principals will not pay any bonus.

Table V in the paper (p.140), describes the bonus-to-effort relation in bonus contracts in two sessions of the experiment (S3 and S4). According to this table, low efforts (between 2 and 4) were made in $35/198 = 17.6\%$ of the cases, suggesting that the percentage of inequity-averse agents is $17.7\%$. Among those agents who made a high effort (of 5 or more), the percentage of those who received a bonus exceeding 21 is $36/127 = 28.3\%$. If the model being tested is correct, the percentage of inequity-averse principals (who pay the bonus) is therefore $28.3\%$. It follows that the estimated percentage of inequity-averse principals is substantially higher (by over 60%) than that of inequity-averse agents, and that both percentages are significantly lower than the 40% required by the

[16] The relevant effort range is $e_1 + e_2 \geq 13$. When $w = 225$ and $e_1 + e_2 \leq 12$, the principal's payoff is lower than the agent's, and so the principal will not pay a bonus.

parametrized theory.

In an unpublished appendix to the paper [18, p.13], the authors expand the scope of Table V by pooling the data of all the bonus contract games of the various treatments (sessions S3 thru S6).[17] The authors say that there is no statistically significant difference in the bonus-effort relation between the two tables (Result 6(b), p.142) but we find that the tables differ enough to make it worth recording the difference.

We compute the estimated percentages of inequity-averse principals and agents using the pooled data as we did in the case of the published Table V. The percentage of inequity-averse principals consistent with the pooled data is $23.9\%$ $(51/213 = 0.239)$, but the percentage of inequity-averse agents consistent with the data is $19.1\%$ $(72/376 = 0.191)$. That is to say, the pooled data suggests that there are about $25\%$ more inequity-averse principals than inequity-averse agents. This result is again not consistent with the authors' parametrization, which assumes a perfect correlation between $\alpha$ and $\beta$, and hence that the percentages of inequity-averse agents and inequity-averse principals are the same.

How would the agents behave when offered the wage $w = 15$ (the average wage offer in the experiment: Figure 5, p.138), if they believed that about 24% of the principals are inequity averse, as the pooled data suggests?[18] It is easy to compute that $all$ the agents (both the money-maximizing agents with $\alpha = \beta = 0$ and the inequity-averse agents with $\alpha = 2$ and $\beta = 0.6$) would choose effort level 3. However, the data shows that only 17% to 19% chose this effort level. The percentage of those choosing effort levels exceeding 6 is about 50%.

It is a fundamental assumption of the theory that bonuses are paid by inequity-averse principals to equate the payoffs of the agent and the principal. We test whether the data confirms this prediction. To perform this test we need to know the empirical wage paid to the agent. The authors don't provide detailed information about the wage paid; they provide only the average wage as a function of time (figure 5). To test whether the principals attempted to equate payoffs via the bonuses, we take the wage paid to be $w = 15$, which is the theoretical wage, and also happens to be the average empirical wage. We compute the fraction of bonuses which give the agent at least $80\%$ of the principal's share; these are the

[17] In all the sessions, the bonus contract game was played. The sessions differ in their framing, and in the number of contract types from which the principals may select.

[18] The result is the same if we take the percentage of inequity-averse principals to be 28% as suggested by Table V.

bonuses for which $b + w \geq (0.8v + c)/1.8$.

We consider only bonuses of 6 or more (assuming that those principals who paid a bonus below 6 are money-maximizers who meant to pay $0$). We find only $44/106 = 41.5\%$ of these bonuses come near to equating the payoffs of the agent and the principal, whereas the theory predicts that all positive bonuses will equate the payoffs. For the pooled data of the unpublished appendix [18, p.13], the fraction of the bonuses that roughly equate the payoffs is $43.7\%$. Clearly, the bonuses paid in the experiment are not all intended to equate the payoffs of the principal and the agent.

The authors claim that the bonuses paid in the experiment form a substantial part of the agent's compensation. They calculate the average bonus (10.4) paid in the experiment, and the average wage (15). The bonus part of the total compensation to the agent is: $10.4/25.4 = 40.9\%$ (Result 5, p.137). We now compute this ratio for the equilibrium (of the 40-60 parametrization). The average bonus is $25\times0.4\times0.6+1\times0.4\times0.4 = 6.16$, and the wage is $15$. Hence the theoretical bonus part of the agent's total compensation is: $6.16/(15 + 6.16) = 29.1\%$. This is much lower than the experimental value of $40.9\%$. The subjects in the experiment therefore cooperate more than the theory permits.

We briefly compare the equilibrium with the experimental behavior under an incentive contract. According to the theory (Proposition S2 of [18, p.2]), all principals demand an effort level of $e^* = 4$, the money-maximizing principals offer a wage of $w = 4$, and the inequity-averse principals offer a wage of $w = 17$. All agents accept the inequity-averse offer and inequity-averse agents reject the money-maximizing offer. Comparing this with the data (the first part of Table III, p. 132), we find that the percentage of principals who offer a wage between 10 and 20 is $26/56 = 46.4\%$. If the theory were right, the percentage of inequity-averse principals would therefore also need to be $46.4\%$. The percentage of inequity-averse agents (those who reject a low wage offer) would need to be $8/26 = 30.7\%$, which is substantially lower than the fraction of inequity-averse principals.

Finally, the average payoff of a principal doesn't match the theory. In the experiment, the average payoff is $8.6$. According to the theory, the average payoff is a weighted average of $15.6$ and $13$ : namely, $15.6(1 - q) + 13q$ where $q$ is the fraction of inequity-averse principals. But this expression is well above $8.6$ for all values of $q$.

### A.3 Fehr, Kremhelmer and Schmidt, 2005 [20]

For this game, the theory predicts that a money-maximizing $B$ player exerts effort $b = 10$, and an inequity-averse player sets $b = 1$ (Proposition 3, p.25 and Proposition 4, p.26). A money-maximizing $A$ player exerts effort $a = 1$, and an inequity-averse $A$ player sets $a = b$.

According to Table 3 on p.17, the fraction of $B$ agents who chose low effort levels 1,2, or 3 is $10/187 = 5.34\%$, and so the percentage of inequity-averse agents would also need to be $5.34\%$.

Of those $B$ players who expended high efforts $(b = 8, 9, or 10)$, a fraction $108/155 = 69.67\%$ were rewarded by a high effort from the $A$ player, and so the percentage of inequity-averse principals would also need to be $69.67\%$.

The huge discrepancy between the percentages of inequity-averse agents and inequity-averse principals will be apparent. Such a discrepancy is incompatible with the authors' parametrization. Moreover, such a high percentage of inequity-averse principals $(69.67\%)$ cannot support the equilibrium considered by authors; if the percentage of inequity-averse principals is higher than 41.2, then some inequity-averse agents (depending on their $\alpha$ values) may cooperate.

If we assume that there are $69.67\%$ inequity-averse individuals with high $\alpha$ *and* $\beta$, it may be possible to explain the data by assuming that the high percentage of inequity-averse principals induced most of the inequity-averse agents to cooperate, leaving behind the $5.34\%$ who shirked. Those who shirked must have a very large $\alpha$ to make them reject the nearly certain bonus. All the other inequity-averse agents must have a lower $\alpha$, which induces them to cooperate. Such an explanation amounts to assuming a particular distribution of the $\alpha$ within the inequity-averse group of agents that depends on the percentage of inequity-averse principals.

# References


[1] R. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6–19, 1995.

[2] Robert J. Aumann. Irrationality in game theory. In *Collected Papers of Robert J. Aumann, Volume I*, pages 621–634. MIT Press, Cambridge, Massachusetts, 2000.

[3] S. Ball and C. Eckel. Buying status: Experimental evidence on status in negotiation. *Psychology and Marketing*, 105:381–405, 1996.

[4] K. Binmore. *Playing Fair: Game Theory and the Social Contract I*. MIT Press, Cambridge, MA, 1994.

[5] K. Binmore. *Just Playing: Game Theory and the Social Contract II*. MIT Press, Cambridge, MA, 1998.

[6] K. Binmore. Economic man—or straw man? a commentary on Henrich et al. *Behavioral and Brain Science*, 28:817–818, 2005.

[7] K. Binmore. *Natural Justice*. Oxford University Press, New York, 2005.

[8] K. Binmore. *Does Game Theory Work? The Bargaining Challenge*. MIT Press, Boston, 2007.

[9] K. Binmore, J. Gale, and L. Samuelson. Learning to be imperfect: The Ultimatum Game. *Games and Economic Behavior*, 8:56–90, 1995.

[10] K. Binmore, J. McCarthy, G. Ponti, A. Shaked, and L. Samuelson. A backward induction experiment. *Journal of Economic Theory*, 184:48–88, 2002.

[11] G. Bolton. A comparative model of bargaining: Theory and evidence. *American Economic Review*, 81:1096–1136, 1991.

[12] G. Bolton and A. Ockenfels. A theory of equity, reciprocity and competition. *American Economic Review*, 90:166–193, 2000.

[13] C. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, Princeton, NJ, 2003.

[14] C. Camerer and D. Harless. The predictive utilty of generalized expected utility theories. *Econometrica*, 62:1251–1290, 1994.

[15] C. Camerer, E. Johnson, T. Rymon, and S. Sen. Cognition and framing in sequential bargaining for gains and losses. In A. Kirman K. Binmore and P. Tani, editors, *Frontiers of Game Theory*. MIT Press, Cambridge, MA, 1994.

[16] M. Deutsch. *Distributive Justice: A Social Psychological Perspective*. Yale University Press, Newhaven, 1985.

[17] E. Fehr and S. Gächter. Cooperation and punishment in public goods experiments. *American Economic Review*, 90:980–994, 2000.

[18] E. Fehr, A. Klein, and K. Schmidt. Appendix to fairness and contract design, 2007. See http://www.econometricsociety.org/ecta/supmat/ECTA5182SUPP.pdf

[19] E. Fehr, A. Klein, and K. Schmidt. Fairness and contract design. *Econometrica*, 114:121–154, 2007.

[20] E. Fehr, S. Kremhelmer, and K. Schmidt. Fairness and optimal allocation of property rights. Discussion Paper 5369, CEPR, London, 2005.

[21] E. Fehr and K. Schmidt. A theory of fairness, competition and cooperation. *Quarterly Journal of Economics*, 114:817–868, 1999.

[22] E. Fehr and K. Schmidt. Theories of fairness and reciprocity: Evidence and economic applications. In S. Dewatripont and L. Hansen, editors, *Advances in Economic Theory: Eighth World Congress (Volume I)*, pages 208–257. Cambridge University Press, Cambridge, 2003.

[23] E. Fehr and K. Schmidt. Fairness and incentives in a multi-task principal-agent model. *Scandinavian Journal of Economics*, 106:453–474, 2004.

[24] E. Fehr and K. Schmidt. The role of equality, efficiency,and rawlsian motives in social preferences. Working Paper 179, University of Zurich, 2004.

[25] E. Fehr and K. Schmidt. Theoretical appendix to fairness and incentives in a multi-task principal-agent model. See http://www.vwl.uni-muenchen.de/ls_schmidt/experiments/multi_task/index.htm, 2004.

[26] E. Fehr and K. Schmidt. The rhetoric of inequity aversion—a reply. See http://www.najecon.org/naj/cache/666156000000000616.pdf, 2005.

[27] E. Fehr and K. Schmidt. The economics of fairness, reciprocity and altruism - Experimental Evidence and new theories. In S. Kolm and J. Ythier, editors, *Handbook of the Economics of Giving, Altruism and Reciprocity*, page 615, North Holland Publishing Co., Amsterdam, 2006.

[28] U. Fischbacher, C. Fong, and E. Fehr. Fairness, error and the power of competition. Working Paper 133, University of Zurich, 2003.

[29] M. Friedman. The methodology of positive economics. In M. Friedman, editor, *Essays on Positive Economics*. Chicago University Press, Chicago, 1953.

[30] D. Gode and S. Sunder. Zero-intelligence traders: Market as a partial sustitute for individual rationality. *Journal of Political Economy*, 101:119–137, 1995.

[31] F. Guala. *The Methodology of Experimental Economics*. cambridge University Press, Cambridge, 2005.

[32] W. Güth, N. Marchand, and J-L. Rulliere. Ultimatum bargaining behavior—a survey and comparison of experimental results. Discussion paper, Humboldt University, Berlin, 1997.

[33] W. Güth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, 3:367–388, 1982.

[34] L. Hansan and J. Heckman. The empirical foundations of calibration. *Journal of Economic Perspectives*, 10:87–104, 1996.

[35] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis *et al*. *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford University Press, New York, 2004.

[36] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis *et al*. "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies *Behavioral and Brain Sciences*, 28: 795–815, 2005.

[37] J. Hey and C. Orme. Investigating generalizations of expected utility theory using experimental data. *Econometrica*, 62:1251–1290, 1994.

[38] C. Hitchens. *The Missionary Position: Mother Teresa in Theory and Practice: Ideology of Mother Teresa*. Verso Books, London, 2003.

[39] E. Hoffman, K. McCabe, K. Sachat, and V. Smith. Preferences, property rights and anonymity in bargaining games. *Games and Economic Behavior*, 7:346–380, 1994.

[40] G. Homans. *Social Behavior: Its Elementary Forms*. Harcourt, Brace and World, New York, 1961.

[41] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–291, 1979.

[42] E. Kayser, T. Schwinger, and R. Cohen. Layperson's conceptions of social relationships: A test of contract theory. *Journal of Social and Personal Relationships*, 1:433–548, 1984.

[43] J. Konow. A positive theory of economic fairness. *Journal of Economic Behavior and Organization*, 31:13–35, 1996.

[44] D. Kreps, P. Milgrom, J. Roberts, and R. Wilson. Rational cooperation in the finitely repeated Prisoners' Dilemma. *Journal of Economic Theory*, 27:245–252, 1982.

[45] J. Ledyard. Public goods: A survey of experimental research. In J. Kagel and A. Roth, editors, *Handbook of Experimental Economics*. Princeton University Press, Princeton, 1995.

[46] M. Lerner. The justice motive in human relations: Some thoughts about what we need to know about justice. In M. Lerner and S. Lerner, editors, *The Justice Motive In Social Behavior*. Plenum, New York, 1981.

[47] M. Lerner. Integrating societal and psychological rules of entitlement: The basic task of each social actor and a fundamental problem for the social sciences. In R. Vermunt and H. Steensa, editors, *Social Justice in Human Relations I: Societal and Psychological Origins of Justice*. Plenum, New York, 1991.

[48] D. Levine. Modeling altruism and spite in experiments. *Review of Economic Dynamics*, 1:593–622, 1998.

[49] C. Manski. Identification of decision rules in experiments on simple games of proposal and response. *European Economic Review*, 46:880–891, 2002.

[50] M. Mitzkewitz and R. Nagel. Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory*, 22:171–198, 1993.

[51] E. Ostrom, J. Walker, and R. Gardner. Covenants with and without the sword: Self-governance is possible. *American Political Science Review*, 86:404–417, 1992.

[52] H. Reis. The mutidimensionality of justice. In R. Folger, editor, *The Sense of Injustice: Social Psychological Perspectives*. Plenum, New York, 1984.

[53] A. Roth, V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review*, 81:1068–1095, 1991.

[54] D. Sally. Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7:58–92, 1995.

[55] E. Sampson. On justice as equality. *Journal of Social Issues*, 31:54–64, 1975.

[56] L. Samuelson. Does evolution eliminate dominated strategies? In A. Kirman K. Binmore and P. Tani, editors, *The Frontiers of Game Theory*. MIT Press, Cambridge, MA, 1994.

[57] L. Samuelson. Foundations of human sociality: A review essay. To appear in *Journal of Economic Literature.*, 2005.

[58] D. Schmidt and T. Neugebauer. Testing expected utility in the presence of errors. *Economic Journal*, 117:470–485, 2007.

[59] S. Schwartz. The justice of need and the activation of humanitarian norms. *Journal of Social Issues*, 31:11–136, 1975.

[60] R. Selten. The equity principle in economic behavior. In H. Gottinger and W. Leinfellner, editors, *Decision Theory and Social Ethics, Issues in Social Choice*. Reidel, Dordrecht, Netherlands, 1978.

[61] R. Selten and R. Stocker. End behavior in finite sequences of prisoners' dilemma supergames: A learning theory approach. *Journal of Economic Behavior and Organization*, 7:47–70, 1986.

[62] J. Steiner. A trace of anger is enough: On the enforcement of social norms. *Economics Bulletin*,8:1–4, 2007.

[63] A. Tversky. *Preference, Belief, and Similarity*. MIT Press, Cambridge MA, 2003.

[64] G. Wagstaff. *An Integrated Psychological and Philosophical Approach to Justice*. Edwin Mellen Press, Lampeter, Wales, 2001.

[65] E. Walster, E. Berscheid, and G. Walster. New directions in equity research. *Journal of Personality and Social Psychology*, 25:151–176, 1973.

[66] E. Walster and G. Walster. Equity and social justice. *Journal of Social Issues*, 31:21–43, 1975.

[67] T. Yamagishi. The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51:110–116, 1986.