Journal of Economic Behavior & Organization 73 (2010) 87–100

Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

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

Experimental economics: Where next?

Ken Binmore [a,*], Avner Shaked [b]

[a] Economics Dept, University College, Gower Street, London WC1E 6BT, United Kingdom
[b] Economics Dept, Bonn University, Adenauerallee 24, 53113 Bonn, Germany

ARTICLE INFO

Article history:
Received 14 November 2007
Accepted 8 October 2008

JEL classification:
C72
C81
C92

ABSTRACT

Where should experimental economics go next? This paper uses the literature on inequity aversion as a case study in suggesting that we could profit from tightening up our act.

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 follow those experimental economists who seek recognition of their subject as a science by adopting the scientific standards that operate in neighboring disciplines like biology or psychology? Or should we follow the tradition in policy-orientated economics of treating experimental results as just another 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 break with the less respectable traditions of economic debate, and join the rest of the scientific community in adopting a more skeptical attitude when far-reaching claims about human behavior are extrapolated from very slender data. As a case study, we examine some of Fehr and Schmidt's influential papers on inequity aversion (Fehr and Schmidt, 1999, 2003, 2004a; Fehr et al., 2005, 2007a). The question is whether the claims made on behalf of the theory are actually supported by the quoted data.

The fact that we chose the theory of inequity aversion for our case study should not be interpreted as a disavowal of this or any other behavioral theory featuring social preferences. We acknowledge that the accumulation of experimental evidence can be regarded as an informal proof that such preferences exist. Nor are we concerned with personalities. We could have written a paper that documented similar interpretive problems in the work of a selection of other authors regarded as being at the top of their profession, both within experimental economics and elsewhere. Fehr and Schmidt themselves are widely admired economists. Fehr is a past president of the Economic Science Association and Schmidt is a member of the Berlin-Brandenburg Academy of Sciences. However, we chose not to write such a wide-ranging critique because our specific claims would then have had to be taken on trust rather than being closely documented, as in the current paper.

* Corresponding author.
E-mail address: k.binmore@ucl.ac.uk (K. Binmore).

0167-2681/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jebo.2008.10.019

2. The optimizing paradigm

This section tries to set the record straight on a number of relevant issues that are sometimes overlooked.

De gustibus non est disputandum. Behavioral economists sometimes claim that neoclassical economists hold that people are selfish. Henrich et al. (2004) go so far as to assert the existence of a "selfishness axiom".[1]

But no such axiom appears in standard economics textbooks. On the contrary, economists are traditionally 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 do not 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.

So behavioral and neoclassical economists have a joint interest in resisting those members of our profession who actually do believe and teach that postulating selfishness is a necessary tenet of economic theory. Reports that students majoring in economics are more likely to exhibit selfish attitudes may be a sign of how widespread this misconception is. On the other hand, theory would seem to place no constraint on the extent to which different empirical schools may legitimately disagree on the economic contexts within which attributing selfishness to economic agents is an adequate approximation to actual behavior.

Maximizing money? To say that agents are money-maximizers does not imply that they are selfish. If Mother Teresa (Hitchens, 2003) 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.

Maximizing money in games. Although neoclassical economics offers no theoretical support for identifying utils with dollars, there is actually much experimental evidence from laboratory games that supports the practice in certain circumstances. All experimental economists accept this claim for market games, but it also holds for most games with money payoffs that have a unique Nash equilibrium with no alternative best replies—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 *Behavioral Game Theory* (2003, p. 46) says that the surveys of Ledyard (1995) and Sally (1995) are too well-known for the evidence to require review.

It does not 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. How else are charitable contributions to be explained? 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, it is well known that there are games in which theoretical predictions based on money payoffs are not robust to small perturbations in the rationality of the players or their utility functions (Akerlof and Yellen, 1985). The finitely repeated Prisoner's Dilemma is one example (Kreps et al., 1982). Public goods games with punishment are another (Steiner, 2007). 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 subjects who are inexperienced or offered an inadequate incentive 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. (2005, p. 801) say of their (inexperienced) subjects: "Experimental play often reflects patterns of interaction found in everyday life." If this is right, then experimentalists need to borrow ideas from 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. However, as in Henrich et al. (2004), it is common in behavioral economics to proceed as though the optimizing paradigm implies backward induction. For example, it is often thought to be unproblematic that the final stage of a finitely repeated game can be treated as though it were a one-shot game. This insistence on backward induction 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.

Bob Aumann (1995) proves that common knowledge of rationality implies backward induction in finite games of perfect information, but his (controversial) definition of rationality is not simply maximizing utility. Nor does Aumann (2000)

[1] We quote from Henrich et al. (2004) here and elsewhere, because the book's long list of coauthors includes a representative jury of prominent members of the behavioral school. For comments, see Samuelson (2005b).

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 encouraged Werner Güth (Güth et al., 1982) to carry out the very first experiment on the Ultimatum Game.

Most game theorists are equally skeptical of backward induction, having abandoned this and other refinement theories twenty years ago. Their skepticism deepened after it was discovered that simple trial-and-error adjustment processes may easily take players in a game to a Nash equilibrium that is not subgame-perfect, or which is weakly dominated (Samuelson, 1994). This is true, for example, of the replicator dynamics in the Ultimatum Game (Binmore et al., 1995).

Of the very large numbers of experimental papers that refute backward induction in the laboratory, our favorite is Camerer et al. (1994), reviewed in Camerer (2003). In three-stage Ultimatum Games with alternating offers, subjects sometimes do not 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, 2007). In a recent paper, we show that even attributing other-regarding utility functions to the subjects that take account of both players' money payoffs cannot rescue backward induction in two-stage Ultimatum Games, regardless of the functional form of the utility functions (Binmore et al., 2002). But like the zombies in horror movies that keep getting up no matter how many bullets arepumped into them, backward induction seemingly cannot be laid to rest.

3. Testing theories scientifically

How should a theory of human behavior be tested in the laboratory? We do not 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, 2005). One might, for example, seek to emulate the psychological work of Tversky (2003). 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 would not 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 do not always succeed in measuring up to our aspirations.

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 (2002) 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 was used neither in formulating the theory nor 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 cannot do this with macroeconomic field data, but the data gathered in economic laboratories is not 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.

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, if a theory predicts that a certain function is exponential, then it also predicts that the function is monotonic. However, data that shows the function is indeed monotonic does not support the theory if it also shows that the function is logarithmic.

Favorable selection. The previous example also serves to illustrate the obvious point that some events are easier to predict than others. A good theory will successfully predict events that are difficult to predict.

Choosing the events that a theory is to predict in a manner that favors the theory—especially if done after the test experiment has been run—is 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. Selecting the rival theories favorably is no more acceptable than selecting the events to be predicted in a favorable manner.

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 that is supposedly being predicted. Hansan and Heckman (1996) is the appropriate authority.

The history of non-expected utility theory provides a good example. Kahneman and Tversky (1979) showed that Von Neumann and Morgenstern's theory of expected utility is a bad predictor in the laboratory. So various alternative theories

were proposed that fitted the data better than expected utility theory when their additional parameters were suitably chosen. This work generated much enthusiasm, and many applied papers were written incorporating one or another non-expected utility theory. But we now have two authoritative papers in the same issue of *Econometrica* showing that, when like is compared with like, all extant theories predict badly—but orthodox expected utility theory arguably performs as well as any rival (see Camerer and Harless, 1994; Hey and Orme, 1994; also a recent paper by Schmidt and Neugebauer, 2007).

Straightforward 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 frankly acknowledged. The methodology should be clearly explained with a view to assisting replication. Failure to report pilot studies is not good practice. Relegating significant information to footnotes or technical appendices should be avoided.

Serious refereeing. Our claims in this paper have been checked down to the last rounding error by a referee who took his duty very seriously. If the same had been true of the referees of the papers examined in the rest of this paper, we would have no grounds for complaint.

4. Inequity aversion

We believe that the work of Fehr and Schmidt on the theory of inequity aversion is appropriate as a case study because it has achieved iconic status within the economics profession. We cannot discuss all their work, which appears in major peer-reviewed journals (including *Econometrica*), and so focus on the foundation stone of their research program, which is a hugely influential paper that appeared in the *Quarterly Journal of Economics* (Fehr and Schmidt, 1999). However, since the *QJE* paper refers to the results of several further papers, there remains much ground to cover. Section 9 also considers other papers, namely Fehr and Schmidt (2004a) and Fehr et al. (2007a, 2005).

At our last count, Google Scholar reported 1831 citations of the *QJE* paper, and 559 citations of the remaining articles discussed in this paper. The number of citations increases day-by-day.

On our behalf, Esshan Vallizadeh examined the content of the works in which the citations appear. Some are experimental papers that offer new data for or against inequity aversion as a predictive tool. Some seek to apply inequity aversion to new situations. A few question the adequacy of Fehr and Schmidt's data or the range of application of their model, suggesting that it is too one-dimensional to seek to explain human behavior on the basis of payoff differences alone. Examples are Charness and Rabin (2002) and Engelmann and Strobel (2004). However, the vast majority of papers simply quote Fehr and Schmidt's work as a leading example of a successful research program in behavioral economics. Some widely cited references are Deaton (2003, p. 27) cited 309 times, Tirole (2002, p. 637) cited 58 times, Huck et al. (2001, p. 11) cited 58 times, and Samuelson (2005a, p. 70) cited 44 times. The work is thought so unproblematic that it is even taught to undergraduates (Wilkinson, 2008).

The significant feature of this literature is that not one of the 2390 works citing Fehr and Schmidt on inequity aversion offers anything but uncritical acceptance of their methods. We found no papers at all (other than our own) that offer any kind of critique of Fehr and Schmidt's methodology, but we hope that the rest of this paper will convince the reader that any close examination of their claims (let alone a serious attempt at replication) would surely have given rise to at least some doubts.

The theory of inequity aversion. The idea of fitting a utility function incorporating inequity aversion to experimental data originates with Gary Bolton (1991), who later refined the theory jointly with Ockenfels (Bolton and Ockenfels, 2000). 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) 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 (α_i, β_i). The parameter α_i measures player i's envy at being poorer than others. The parameter β_i measures player i's discomfort at being richer. Fehr and Schmidt also follow Levine (1998) in allowing players to be heterogeneous, so that a population is characterized by a joint distribution of α and β.

As the subject population in the theory of Fehr and Schmidt is heterogeneous, 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 α and β, 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 et al. (1994) show that subjects who believe that they have earned the right to propose in the Ultimatum Game expect more. Ball and Eckel (1996) 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 claimed by the proposer (Mitzkewitz and Nagel, 1993). The anthropological studies reported in Henrich et al. (2004) 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 (1996) is the most recent economist to press this point, but Selten (1978) drew our attention to the work of Homans (1961) many years ago. It is something of a scandal that we ignore a substantial experimental literature[2] in social psychology, which claims 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 situations in which fairness attitudes matter. We recognize that new theories sometimes have a wider domain of application than their authors explore,[3] but we choose only to examine applications proposed by Fehr and Schmidt themselves.

5. Fehr and Schmidt's *QJE* article

In a survey paper prepared for an invited lecture at the World Congress of 2000, Fehr and Schmidt (2003, p. 222) summarize the work that stems from their *QJE* paper in the following terms:

> Using data that are available from many experiments on the [Ultimatum Game], Fehr and Schmidt calibrate the distribution of α and β 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 Section 3. However, we now address several problems with the claims they make in this passage. In particular:

1. The Ultimatum Game data that is said to have been used to calibrate the parameters α and β is inadequate for this purpose. It is logically impossible to tie down the parameters from the data said to have been 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 do not keep the distribution of parameters constant. This is a major cause for concern, 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 do not 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 favor their theory both in what they choose to regard as a prediction, and in the variant of a money-maximizing model with which they choose to compare their model.
4. In a reply to an earlier critique on Shaked's website (see Shaked, 2005), Fehr and Schmidt (2005) 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.[4] 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, p. 222) say they used data from Ultimatum Game experiments to calibrate their inequity-aversion model in their *QJE* paper. The resulting marginal and joint distributions they say they later keep fixed 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 α, and that of proposers by their discomfort parameter β. So the marginal distribution of β needs to be calculated from data on the proposers' offers, while the marginal distribution of α needs to be calculated from data on the responders' acceptance rates.

However, if the inequity aversion theory holds in the Ultimatum Game, no more can be said of proposers who offered 50% of the available money than that their β exceeds 0.5 (Fehr and Schmidt, 1999, Proposition 1, p. 826). But it will be noted

[2] The literature offers experimental support for Aristotle's contention that what is fair is what is proportional. For example, Deutsch (1985); Kayser et al. (1984); Lerner (1981, 1991); Reis (1984); Sampson (1975); Schwartz (1975); Wagstaff (2001); Walster et al. (1973), and Walster and Walster (1975).

[3] Max Planck's initial use of quanta to fit data on black body radiation is a famous example.

[4] Their reply also cites two further papers that are not mentioned in their invited address (Fehr and Schmidt, 2003). We do not discuss Fischbacher et al. (2003) because it depends on quantal response equilibria, whose explanatory value remains controversial. Fehr and Schmidt (2004c) does not seem relevant to their *QJE* parametrization at all.

Table 1
Distributions of α and β 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" (Fehr and Schmidt, 1999, p. 864).

	%
Marginal distribution (α)	
0.0	30
0.5	30
1.0	30
4.00	10
Marginal distribution (β)	
0.0	30
0.25	30
0.6	40
Joint distribution (α, β)	
(0.0, 0.0)	30
(0.5, 0.25)	30
(1.0, 0.6)	30
(4.0, 0.6)	10

that 40% of subjects are assigned a value of $\beta = 0.6$ in Table 1. Fehr and Schmidt (2005, p. 7) defend 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 β_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 β in their model to be 0.55, their model would not have been consistent with the data from the Public Goods Game with Punishment in Fehr and Gächter (2000) of Section 8.2. If they had chosen the highest value of β to be 0.85, their model would not have been consistent with the Competition among Responders Game Güth et al. (1997) of Section 8.3.

The Ultimatum Game data that would be needed to estimate the joint distribution of (α, β) reproduced in Table 1 seems to be absent altogether. For this purpose, we would 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 α and β, they presumably have another source in mind (Fehr and Schmidt, 1999, p. 864).

Thus it appears that Fehr and Schmidt did not estimate their parameters from Ultimatum Game data alone, which leaves the parameters under-identified, and able to float to a considerable degree. It would be useful for someone to estimate the most likely range of the parameters econometrically, and we hope this will be done to clarify the situation.

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, 2000). Fehr and Schmidt's analysis of this game is based on their Proposition 5, the proof of which appears at the very end of their appendix. It is only then that we learn of the introduction of the joint distribution of α and β that appears in our Table 1. However, the proof of Proposition 5 would seem to assume that all the types who are not 'conditionally cooperative enforcers' (with $\beta \geq 0.6$ and a correspondingly high α) 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 is the first indication that Fehr and Schmidt sometimes alter their assumptions about the joint distribution of α and β 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 engaged in a *predicting* exercise but a *fitting* exercise, in which parameters can be changed 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 Game without punishment

Section 2 mentions Camerer's (2003) endorsement of the conclusions reached by Ledyard (1995) and Sally (1995) from their surveys of a 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, 1995, p. 172).

It is therefore surprising that Fehr and Schmidt (1999, 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 quoting Ledyard's data to this end for inexperienced subjects, but 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 data rather well.

However, rather than quote the conclusions of Ledyard's comprehensive survey of papers published before 1995, Fehr and Schmidt (1999, p. 838) choose for themselves what experimental studies on Public Goods Games to report in their Table II. A particular feature of their choice to which we wish to draw attention is that this 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 (Fehr and Schmidt, 1999, footnote 18).[5] However, many experiments suggest that laboratory subjects tend to play more cooperatively in the partner design, even when it is common knowledge that only a fixed number of games are to be played (Selten and Stocker, 1986).[6]

It is true that if backward induction had not been overwhelmingly refuted in the laboratory, one could treat the final game in the partner design as though it were a one-shot game, but backward induction has been overwhelmingly refuted in the laboratory—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 (2000), which is a key source of data for Fehr and Schmidt's *QJE* paper. The following table (based on Figs. 2 and 4 of Fehr and Gächter, 2000) shows that no econometrics is necessary to check that the partner design yields different results from the stranger design for experienced subjects:

	Stranger design	Partner design
Without punishment	75% free riders	51% free riders
With punishment	10% contributors	80% contributors

Eyeballing the relevant figures in Fehr and Gächter (2000) is similarly all that is necessary to confirm that inexperienced subjects do not always play in the same way as experienced subjects.

In their *QJE* paper, Fehr and Schmidt could have chosen to predict at least four pieces of data from Fehr and Gächter (2000), depending on whether they chose the stranger or partner design, and whether they chose experienced or inexperienced subjects.[7] In the game with punishment, we shall shortly see that Fehr and Schmidt favor their theory when choosing what to predict by taking the case of the partner design with experienced subjects. If they had chosen to predict the same case in the game without punishment that we are currently considering, 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 do not 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 does not appear in the paper of Fehr and Gächter (2000) at all. They compare the prediction of their theory (nearly 100% free riders[9]) with the average percentage of 73% obtained from their Table II (which summarizes data from a selection of other experiments). Despite this gap between their prediction and the data being predicted, Fehr and Schmidt (1999, p. 845) say:

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

[5] Fehr and Schmidt (1999, Table II) say 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.

[6] A referee asks that we quote Andreoni (1988) as an exception.

[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 show clear differences.

[8] The percentage of 51% is estimated from Fig. II of the Q JE article, which reproduces Fig. 4 in Fehr and Gächter (2000). A precise figure does not seem to be given.

[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%.

A footnote is appended that seeks to explain away the difference of 25% (which is a relative difference of $(98.48 - 73)/73 = 35\%$) 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.

Summary. Fehr and Schmidt (1999) 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 Fig. II (Fehr and Gächter, 2000, Fig. 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 35% 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 (Fig. II in the *QJE* paper) as they do for games with punishment, the gap would have been about 49%. Finally, it is problematic to be considering the partner design at all, when it is standard practice to use the stranger design—especially when they do not make it clear that they are deviating from standard practice.

None of the doubts that this discussion raises seem necessary to us. 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.

8.2. Public Goods Game with punishment

This game is the only case analyzed in the *QJE* paper in which the question is not 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 (2000) show that Yamagishi's (1986) finding that free riders do get punished extends even to the case of one-shot games, 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 methods when arguing in favor of their parametrized model of inequity aversion. We have already mentioned that a value of β 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 α and β.

It is unfortunate that the data selected for prediction should be so arbitrary. Standard practice[10] would call on them to predict the data from the stranger design given in Fig. 2 of Fehr and Gächter (2000) but they choose instead to predict the data from the partner design given in Fig. 4 (reproduced as Fig. 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% contribute the maximum in the stranger design. The resulting confusion would seem unnecessary, since our impression is that the data from the stranger design would still present a major challenge to what Fehr and Schmidt 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 at least three reasons. The first is that treating the final round in the partner design as though it were a one-shot game is to take for granted the validity of backward induction. The second reason 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.

To understand the third reason, first note that Fehr and Gächter's (2000) results for their fixed-horizon game are consistent with those of Ostrom et al. (1992) 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., 1982) 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 and Stocker (1986) are not alone in finding 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

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

Goods Game with Punishment, Steiner (2007) 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. Electing to predict the data from the partner design is therefore only an effective tactic for Fehr and Schmidt if they are also allowed to make a favorable selection of the rival money-maximizing prediction as one which denies that a *small* fraction of the subject population will do something other than maximize money. But who 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 chosen 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 all patterns 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 of one of these equilibria. However, Fehr and Schmidt (1999, p. 842) make a favorable selection of 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.

Summary. It is not clear how Fehr and Schmidt predetermined their floating parameters in a manner that turned out to fit the Public Goods Game with Punishment. They make a favorable selection of 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 that 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 backward induction or the laboratory evidence that militates against it (Binmore et al., 2002).

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 an 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, even modeling the subjects in laboratory markets as 'zero-intelligence' traders—who honor their budget constraint, but otherwise bid at random—is quite successful in predicting final outcomes (Gode and Sunder, 1995).

Market with competition among proposers. The predicted data comes from a paper by Roth et al. (1991). 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. (1997). 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 by an orthodox competitive analysis.

Fehr and Schmidt's Proposition 3 shows that their model has a multiplicity of subgame-perfect equilibria of which one generates the same prediction as the money-maximizing model provided that $\beta < 5/6$. Since the maximum value of β

in their parametrization is 0.6, this requirement is easily accommodated. However, nothing in the Ultimatum Game data restricts the maximum value of β.

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

Fehr and Schmidt (2005) say that it was the results of the three contract papers examined in this section that were decisive in their deciding to make the far-reaching claims quoted in Section 5. However, we find that the papers do not support their claims.

The three papers have the common feature that a principal offers one of various possible contracts to an agent. The agent can then exert a costly effort, which generates a payoff for the principal. All use the stranger design rather than the more problematic partner design. More experiments are reported, but we concentrate on those in which players choose between:

1. Bonus, trust, and incentive contracts (Fehr et al., 2007a). 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 initially owns the whole project but can transfer half the ownership rights to the other player (Fehr et al., 2005).
3. Piecewise and bonus contracts (Fehr and Schmidt, 2004a).

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.

Refinement theory. A major problem of interpretation arises in all three contract papers. It is best illustrated in the first and third papers listed above, because the models that are employed to explain the data are signalling games, which notoriously have many equilibria. In order for a precise prediction to be made, it is necessary to propose some equilibrium refinement that attributes appropriate beliefs to the players in the event of out-of-equilibrium play. The downfall of refinement theory some twenty years ago is often traced to disputes about which out-of-equilibrium beliefs can reasonably be regarded as plausible in such signalling games. However, it is not only the application of a discredited theory that creates difficulties in interpreting the empirical results.

The out-of-equilibrium beliefs assumed in the papers are not determined by the theory of inequity aversion, but appear as auxiliary assumptions that do not seem to be directly testable. The papers only specify these auxiliary assumptions to the extent necessary to tie down a particular Nash equilibrium, which then serves as the prediction of the model. However, most subjects do not play according to this prediction, but deviate to a greater or lesser extent. How does one interpret such out-of-equilibrium play? Does one follow the papers in ignoring the effect that their model says that such deviations will have on the beliefs (and hence the behavior) of other subjects. If not, how does one predict what the change of behavior of other subjects will be without a precise specification of the out-of-equilibrium beliefs maintained in the model? This is one of several reasons why we think the models in these papers are not strictly testable. However, we put these problems aside in what follows.

Learning. The fact that much other work shows that game theory seldom succeeds in predicting the behavior of inexperienced subjects is largely neglected. Even if Fehr and Schmidt's parametrized theory of inequity aversion were correct, it would therefore be surprising to see it instantiated by subjects who have not had an adequate opportunity to learn how to play the predicted equilibrium. If it were instantiated, one would have to ask why subjects in the Prisoners' Dilemma have to learn how to optimize, but subjects in the contract games do not.

9.1. Keeping the distribution constant?

The contract papers do not keep the *QJE* parametrization constant as Fehr and Schmidt claim. In all three 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 *Q* JE paper) and 40% inequity-averse individuals who all have the same α and β that both exceed 0.5. We refer this new distribution as a 40–60 distribution. Such a distribution is not consistent with Ultimatum Game data because Fehr and Schmidt eliminate 30% of the types in our 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 (Fehr and Schmidt, 2004a, p. 470; Fehr et al., 2005, p. 22; Fehr et al., 2007a, 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.[11]

In Fehr et al. (2005), and in Fehr et al. (2007a) all the higher values of α are equated to 2, although the value $\alpha = 2$ appears nowhere in our Table 1.

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 we shall see that these qualitative predictions have been chosen to favor the theory, while other more fundamental predictions that are refuted by the data are neglected.

An exception leads the authors to observe that their theory provides "surprisingly accurate quantitative predictions" (Fehr et al., 2007a, 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, but 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 et al. (2007a) unconsciously echo Milton Friedman's (1953) 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 have not been favorably selected. This is not 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. We find little correspondence between such predictions and the reported data. Our detailed findings are documented in an appendix that cannot be included in the published version of the paper, but is to be found at http://www.wiwi.uni-bonn.de/shaked/BS-FS-appendix/. The rest of the paper itemizes some of the inconsistencies between the predictions of the theory and the reported data we have found.

A particularly revealing statistic is the percentage of principals classified as inequity-averse, since the theory says that it is this percentage that determines the behavior of the agents. If the percentage is high, all types of agent should behave cooperatively by expending high effort. If it is low, no types should cooperate. Only in the intermediate range should the two types of agent behave differently (with money-maximizing types working hard and inequity-averse types shirking). In Fehr et al. (2007a), the estimated fraction of inequity-averse principals is about 27% (rather than the predicted 40%). If the 40–60 distribution is replaced by a 27-73 distribution, the theory predicts no cooperation at all (with even money-maximizing agents then choosing low effort).

In Fehr et al. (2005) the estimated percentage of inequity-averse principals is 78.4%. This differs both from the predicted 40% and the 27% observed in Fehr et al. (2007a,b). If we apply the theory with 78.4% inequity-averse principals, all agents should again expend maximum effort, but 13.1% choose a low effort level. If we treat 13.1% as an estimate of the percentage of inequity-averse agents, then the 40–60 model underpredicts the percentage of inequity-averse principals by a wide margin and overestimates the percentage of inequity-averse agents by a similarly wide margin. Such data would seem difficult to reconcile with any model that does not allow a player's type to be a function of both the game being played and the role assigned to a player in that game.

Equally large discrepancies between the data and the predictions of the theory appear in Fehr and Schmidt (2004a,b). The estimated fraction of inequity-averse agents is 15.7% (instead of the predicted 40%). If this is also the fraction of inequity-averse principals and if the 40–60 distribution is replaced by a 15.7–84.3 distribution, there will be no cooperation at all (even agents with $\alpha^* = 2$ then choose high effort).

[11] In all three contract games, types with $\alpha = 0.5$ and $\beta = 0.25$ are eliminated from the distribution and replaced by money-maximizing types. However, an agent of this type has too high a value of α to play like a money-maximizer in the equilibria the authors nominate as the predictions of their theory. Consider, for example, how an agent with $\alpha = 0.5$ and $\beta = 0.25$ would behave if the principal plays according the equilibrium of Fehr et al. (2007a). If the agent plays like the money-maximizing agents with whom he is pooled in the contract papers, then he will choose the effort level $e = 7$. The payoffs before a bonus is paid are then (5, 55). In 40% of the cases, an inequity-averse principal will then pay a bonus of 25, which yields a final payoff of 30. The other 60% of the time, a money-maximizing principal pays no bonus. His envy is then aroused, which results in a final payoff of $5 - 0.5(55 - 5) = -20$. His expected payoff before the type of the principal is revealed is therefore $0.4 \times 30 + 0.6 \times (-20) = 0$. However, the agent gets most by behaving like an inequity-averse agent, who chooses the effort level $e = 3$. The payoffs before the bonus are then (13, 15), so his final payoff is $0.4 \times (13 + 1) + 0.6 \times (13 - 0.5(15 - 13)) = 12.8$.

[12] Recall that Camerer (2003, 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.

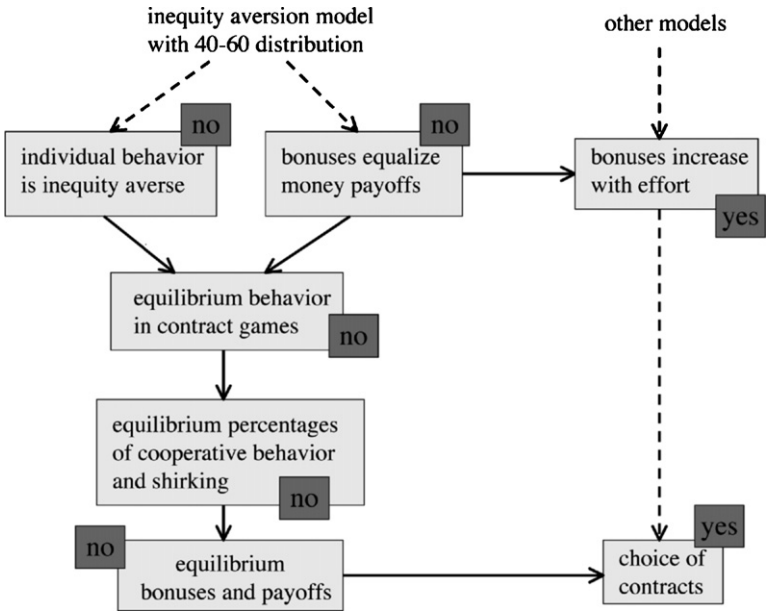

Fig. 1. Neglecting logic. The boxes marked NO indicate predictions of the inequity aversion theory with the 40–60 distribution that the data refutes in the three contract games. The boxes marked YES indicate predictions that are said to be qualitatively accurate. The latter predictions would follow from a variety of models. The firmly drawn arrows indicate the logical structure of the predictions. To claim that the data supports the inequity aversion theory is therefore to neglect to pay attention to the logic of the theory.

A fundamental prediction of the theory is that all inequity-averse principals will pay bonuses that equalize the payoffs of the principal and the agent. (Money-maximizing principals will pay no bonuses at all.) However, we find no support for the claim that all bonuses that are paid tend to equalize the payoffs of the agent and the principal. Fehr et al. (2007a) provides an example. Depending on how much deviation from strict inequality one allows, we find that only between 42% and 62% (instead of 100%) of the sizeable bonuses came close to equalizing the payoffs of the agent and the principal. Fig. 1 sketches the conclusions of our comparison of the data with the theory. The firmly drawn arrows indicate logical 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 to overlook the logic of the theory. The experiments actually point to the need to formulate some new theory, which might perhaps be a reparametrized version of Fehr and Schmidt's inequity-aversion model.

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 (2005a) review of Henrich et al. (2005).)

The research project of Henrich et al. (2005), (of which Fehr is a part) would seem to provide support for this unremarkable possibility. The anthopological studies in the book uncontroversially debunk the idea that we are genetically programmed with culturally independent other-regarding utility functions. Its team of 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.

Summary. Fehr and Schmidt (2003, p. 222) are not entitled to claim that their "calibrated" model yields "quantitatively accurate predictions" in the three contract games, even after altering the *QJE* parametrization to a 40–60 distribution. On the contrary, the data would seem to refute their parametrized model.

10. Conclusion

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 Assisi. We simply recognize that they have different tastes by writing different numbers in their payoff matrices.[13] We are frustrated by the fact that some behavioral economists

[13] Even Binmore's (2005b) theory of fairness norms assumes nothing about whether personal preferences include an other-regarding component or not.

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 persistence of practices that we see as inimical to the future of our profession. It is perhaps inevitable that some readers will misinterpret our paper as an attack on particular authors or theories, but we repeat that our critique is not intended as a refutation of inequity aversion or any other theory of social preference. Nor are our criticisms directed only at some small coterie of authors within experimental economics. On the contrary, we believe they apply more widely within empirical economics than is at all comfortable. Nor do we think that all economic theorists, whether neoclassical or behavioral, are free from sin. Our aim was simply to draw the attention of mainstream economists to the danger of tolerating practices that would be regarded as unscientific in other disciplines.

Acknowledgements

Ken Binmore gratefully acknowledges the financial support of both the British Economic and Social Research Council through the Centre for Economic Learning and Social Evolution (ELSE) and the British Arts and Humanities Research Council through grant AH/F017502. Avner Shaked gratefully acknowledges the financial support of the Deutsche Forschungsgemeinschaft through SFB/TR 15.

References

Akerlof, G., Yellen, J., 1985. Can small deviations from rationality make significant differences in economic equilibria? American Economic Review 75, 708–720.
Andreoni, J., 1988. Why free ride? Strategies and learning in public goods experiments. Journal of Public Economics 37, 291–304.
Aumann, R., 1995. Backward induction and common knowledge of rationality. Games and Economic Behavior 8, 6–19.
Aumann, R.J., 2000. Irrationality in game theory. In: Collected Papers of Robert J. Aumann, vol. I. MIT Press, Cambridge, MA, 621–634.
Ball, S., Eckel, C., 1996. Buying status: experimental evidence on status in negotiation. Psychology and Marketing 105, 381–405.
Binmore, K., 2005a. Economic man—or straw man? A commentary on Henrich et al. Behavioral and Brain Science 28, 817–818.
Binmore, K., 2005b. Natural Justice. Oxford University Press, New York.
Binmore, K., 2007. Does Game Theory Work? The Bargaining Challenge. MIT Press, Cambridge, MA.
Binmore, K., Gale, J., Samuelson, L., 1995. Learning to be imperfect: the Ultimatum Game. Games and Economic Behavior 8, 56–90.
Binmore, K., McCarthy, J., Ponti, G., Shaked, A., Samuelson, L., 2002. A backward induction experiment. Journal of Economic Theory 184, 48–88.
Bolton, G., 1991. A comparative model of bargaining: theory and evidence. American Economic Review 81, 1096–1136.
Bolton, G., Ockenfels, A., 2000. A theory of equity, reciprocity and competition. American Economic Review 90, 166–193.
Camerer, C., 2003. Behavioral Game Theory, Experiments in Strategic Interaction. Princeton University Press, Princeton.
Camerer, C., Harless, D., 1994. The predictive utility of generalized expected utility theories. Econometrica 62, 1251–1290.
Camerer, C., Johnson, E., Rymon, T., Sen, S., 1994. Cognition and framing in sequential bargaining for gains and losses. In: Kirman, A., Binmore, K., Tani, P. (Eds.), Frontiers of Game Theory. MIT Press, Cambridge, MA, pp. 101–120.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Quarterly Journal of Economics 117, 817–869.
Deaton, A., 2003. Health, inequality, and economic development. Journal of Economic Literature 41, 113–158.
Deutsch, M., 1985. Distributive Justice: {A} Social Psychological Perspective. Yale University Press, New Haven.
Engelmann, D., Strobel, M., 2004. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. American Economic Review 94, 857–869.
Fehr, E., Gächter, S., 2000. Cooperation and punishment in public goods experiments. American Economic Review 90, 980–994.
Fehr, E., Klein, A., Schmidt, K., 2007a. Fairness and contract design. Econometrica 114, 121–154.
Fehr, E., Klein, A., Schmidt, K., 2007b. "Fairness and contract design" Supplementary material. See http://www.econometricsociety.org/ecta/supmat/ECTA5182SUPP.pdf.
Fehr, E., Kremhelmer, S., Schmidt, K., 2005. Fairness and optimal allocation of property rights. Discussion Paper 5369. CEPR, London.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition and cooperation. Quarterly Journal of Economics 114, 817–868.
Fehr, E., Schmidt, K., 2003. Theories of fairness and reciprocity: Evidence and economic applications. In: Dewatripont, S., Hansen, L. (Eds.), Advances in Economic Theory: Eighth World Congress, vol. I). Cambridge University Press, Cambridge, pp. 208–257.
Fehr, E., Schmidt, K., 2004a. Fairness and incentives in a multi-task principal-agent model. Scandinavian Journal of Economics 106, 453–474.
Fehr, E., Schmidt, K., 2004b. 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.
Fehr, E., Schmidt, K., 2004c. The role of equality, efficiency, and Rawlsian motives in social preferences. Working Paper 179, University of Zurich.
Fehr, E., Schmidt, K., 2005. The rhetoric of inequity aversion—a reply. See http://www.najecon.org/naj/cache/666156000000000616.pdf.
Fischbacher, U., Fong, C., Fehr, E., 2003. Fairness, error and the power of competition. Working Paper 133, University of Zurich.
Friedman, M., Friedman, M., 1953. The Methodology of Positive Economics. Essays on Positive Economics. Chicago University Press, Chicago, pp. 1–25.
Gode, D., Sunder, S., 1995. Zero-intelligence traders: market as a partial substitute for individual rationality. Journal of Political Economy 101, 119–137.
Guala, F., 2005. The Methodology of Experimental Economics. Cambridge University Press, Cambridge.
Güth, W., Marchand, N., Rulliere, J-L., 1997. Ultimatum Bargaining Behavior—A Survey and Comparison of Experimental Results. Discussion Paper. Humboldt University, Berlin.
Güth, W., Schmittberger, R., Schwarze, B., 1982. An experimental analysis of ultimatum bargaining. Journal of Economic Behavior and Organization 3, 367–388.
Hansan, L., Heckman, J., 1996. The empirical foundations of calibration. Journal of Economic Perspectives 10, 87–104.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., 2004. Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies. Oxford University Press, New York.
Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., 2005. "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. Behavioral and Brain Sciences 28, 795–815.
Hey, J., Orme, C., 1994. Investigating generalizations of expected utility theory using experimental data. Econometrica 62, 1251–1290.
Hitchens, C., 2003. The Missionary Position: Mother Teresa in Theory and Practice: Ideology of Mother Teresa. Verso Books, London.
Hoffman, E., McCabe, K., Sachat, K., Smith, V., 1994. Preferences, property rights and anonymity in bargaining games. Games and Economic Behavior 7, 346–380.
Homans, G., 1961. Social Behavior: Its Elementary Forms. Brace and World, New York, Harcourt.
Huck, S., Mueller, W., Normann, H-T., 2001. Stackelberg beats Cournot: on collusion and efficiency in experimental markets. The Economic Journal 111, 1–17.

Kahneman, D., Tversky, A., 1979. Prospect theory: an analysis of decision under risk. Econometrica 47, 263–291.
Kayser, E., Schwinger, T., Cohen, R., 1984. Layperson's conceptions of social relationships: a test of contract theory. Journal of Social and Personal Relationships 1, 433–548.
Konow, J., 1996. A positive theory of economic fairness. Journal of Economic Behavior and Organization 31, 13–35.
Kreps, D., Milgrom, P., Roberts, J., Wilson, R., 1982. Rational cooperation in the finitely repeated Prisoners' Dilemma. Journal of Economic Theory 27, 245–252.
Ledyard, J., 1995. Public goods: a survey of experimental research. In: Kagel, J., Roth, A. (Eds.), Handbook of Experimental Economics. Princeton University Press, Princeton, pp. 253–279.
Lerner, M., 1981. The justice motive in human relations: some thoughts about what we need to know about justice. In: Lerner, M., Lerner, S. (Eds.), The Justice Motive In Social Behavior. Plenum Press, New York, pp. 211–247.
Lerner, M., 1991. Integrating societal and psychological rules of entitlement: The basic task of each social actor and a fundamental problem for the social sciences. In: Vermunt, R., Steensa, H. (Eds.), Social Justice in Human Relations I: Societal and Psychological Origins of Justice. Plenum Press, New York.
Levine, D., 1998. Modeling altruism and spite in experiments. Review of Economic Dynamics 1, 593–622.
Manski, C., 2002. Identification of decision rules in experiments on simple games of proposal and response. European Economic Review 46, 880–891.
Mitzkewitz, M., Nagel, R., 1993. Experimental results on ultimatum games with incomplete information. International Journal of Game Theory 22, 171–198.
Ostrom, E., Walker, J., Gardner, R., 1992. Covenants with and without the sword: self-governance is possible. American Political Science Review 86, 404–417.
Reis, H., 1984. The mutidimensionality of justice. In: Folger, R. (Ed.), The Sense of Injustice: Social Psychological Perspectives. Plenum Press, New York, pp. 111–131.
Roth, A., Prasnikar, V., Okuno-Fujiwara, M., Zamir, S., 1991. Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: an experimental study. American Economic Review 81, 1068–1095.
Sally, D., 1995. Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. Rationality and Society 7, 58–92.
Sampson, E., 1975. On justice as equality. Journal of Social Issues 31, 54–64.
Samuelson, L., 1994. Does evolution eliminate dominated strategies? In: Kirman, A., Binmore, K., Tani, P. (Eds.), The Frontiers of Game Theory. MIT Press, Cambridge, MA, pp. 175–189.
Samuelson, L., 2005a. Economic theory and experimental economics. Journal of Economic Literature 43, 65–107.
Samuelson, L., 2005b. Foundations of human sociality: a review essay. Journal of Economic Literature 43, 488–497.
Schmidt, D., Neugebauer, T., 2007. Testing expected utility in the presence of errors. Economic Journal 117, 470–485.
Schwartz, S., 1975. The justice of need and the activation of humanitarian norms. Journal of Social Issues 31, 11–136.
Selten, R., 1978. The equity principle in economic behavior. In: Gottinger, H., Leinfellner, W. (Eds.), Decision Theory and Social Ethics, Issues in Social Choice. Reidel, Dordrecht, Netherlands, pp. 289–305.
Selten, R., Stocker, R., 1986. End behavior in finite sequences of prisoners' dilemma supergames: a learning theory approach. Journal of Economic Behavior and Organization 7, 47–70.
Shaked, A., 2005. The rhetoric of inequity aversion. See http://www.najecon.org/naj/cache/666156000000000612.pdf.
Steiner, J., 2007. A trace of anger is enough: on the enforcement of social norms. Economic Buletin 8, 1–4.
Tirole, J., 2002. Rational irrationality: some economics of self-management. European Economic Review 46, 633–655.
Tversky, A., 2003. Preference, Belief, and Similarity. MIT Press, Cambridge MA.
Wagstaff, G., 2001. An Integrated Psychological and Philosophical Approach to Justice. Edwin Mellen Press, Lampeter, Wales.
Walster, E., Berscheid, E., Walster, G., 1973. New directions in equity research. Journal of Personality and Social Psychology 25, 151–176.
Walster, E., Walster, G., 1975. Equity and social justice. Journal of Social Issues 31, 21–43.
Wilkinson, N., 2008. An Introduction to Behavioral Economics. Palgrave Macillan, Basingstoke, UK.
Yamagishi, T., 1986. The provision of a sanctioning system as a public good. Journal of Personality and Social Psychology 51, 110–116.