GOVERNANCE AND THE EFFICIENCY OF ECONOMIC SYSTEMS

GESY

Discussion Paper No. 164

On the Explanatory Value of Inequity Aversion Theory

Avner Shaked*

September 2006

*Avner Shaked, Department of Economics, University of Bonn, Adenauerallee 24, 53113, Bonn, Germany.

Financial support from the Deutsche Forschungsgemeinschaft through SFB/TR 15 is gratefully acknowledged.

On the Explanatory Value of Inequity Aversion Theory

Avner Shaked[1]
Economics Department
Bonn University
Adenauerallee 24
53113, Bonn, Germany

September 20, 2006

[1]Financial support by the DFG (German Science Foundation) under SFB/TR-15 is gratefully acknowledged.
Special thanks to Dirk Engelmann, Ariel Rubinstein and Ran Spiegler for interesting and helpful conversations on methodology.

Abstract

In a number of papers on their theory of Inequity Aversion, E. Fehr and K. Schmidt have claimed that the theory explains the behavior in many experiments.

By virtue of having an infinite number of parameters the theory can predict a wide range of outcomes, from the competitive to the cooperative. Its prediction depends on values of these parameters.

Fehr & Schmidt provide no explicit methodological plan for their project and as a result they repeatedly make logical and methodological errors. We look at the methodology of their explanations and find that no connection has been established between the experimental data and the behavior predicted by the theory. We conclude that the theory of inequity aversion has no explanatory value beyond its trivial capacity to predict a broad range of outcomes as a function of its parameters.

1 Introduction

The theory of Inequity Aversion was introduced by E. Fehr and K. Schmidt in [Fehr and Schmidt 1999]. Its aim was to provide a unified interpretation of seemingly contradictory experimental evidence in one-shot games, which cannot be explained by the traditional assumption of rational selfish agents. The theory of Inequity Aversion asserts that individuals have preferences over the distribution of payoffs in their group and that they are averse to inequity. Except for this departure from the traditional selfish preferences, the theory continues to assume that all individuals act rationally and maximize their utility, given their preferences and their information about the preferences of others.

The theory does not restrict the distribution of inequity aversion in the population, its predictions, therefore, depend on the composition of the relevant population. For any given game the theory's prediction depends on how inequity averse the population is. If all individuals are rather selfish, the prediction will not differ much from that of the traditional theory with selfish preferences, while if many individuals care a great deal about equity, the theory will tend to predict fair, egalitarian outcomes. Thus, there is a broad spectrum of outcomes that are compatible with the theory, depending on how inequity averse the population is.

At this point the theory has little explanatory value since it is compatible with nearly all behaviors. It is obvious that by adding some individuals with a built-in preference for egalitarian allocations the theory can be compatible with experimental behavior that cannot be explained by the traditional selfish model applied to one-shot games. Still, the model could have served as a useful theoretical tool to study the interactions between selfish and inequity averse individuals, but Fehr and Schmidt wanted to achieve much more than that. Fehr and Schmidt intended to improve the explanatory power of their theory by calibrating the model (fixing a population) and using this population to explain the experimental behavior in various games.

The theory has immediately won great popularity among economists, not least because of the authors' claim that the theory can explain many experiments. Google Scholar, the search engine for scholarly literature, lists (in July 2006) over 1200 citations of the paper.

Since their original paper, Fehr & Schmidt continued to write a number of papers in which they use the calibrated model to explain various experiments. The underlying assumption of the theory is that each individual has a fixed degree of inequity aversion. In their latest paper Fehr & Schmidt admit that the behavior of the individuals does not fit the theory, but they continue to use a calibration of the model. Using a calibration of the model, despite the fact that individuals do not follow the theory, amounts to assuming that, rather mysteriously, the population as a whole has a fixed distribution of inequity aversion. However, throughout their papers, Fehr & Schmidt do not keep their calibration fixed.

Fehr & Schmidt prove detailed propositions based on their calibrations, but they do not test these predictions, they only confirm that the theory's final conclusion (choice of contract) matches the data. The theory's detailed predictions are the fundamental and crucial factors from which the final conclusion is derived. A closer look at the theory and the data reveals that the calibrated theory strongly disagrees with the data, that the theory is irrelevant to the data, and

that the agreement of the final conclusions does not show that the population is motivated by inequity aversion considerations. The theory adds nothing to the understanding of the data and has no explanatory value.

In addition, Fehr & Schmidt's process of selecting the calibration is riddled with methodological errors. When they select the calibration (in [Fehr and Schmidt 1999]) Fehr & Schmidt use the data of some of the experiments which they intend to explain. Having selected the calibration, they do not keep it fixed. To fit the data of their 'most important' experiment, the only experiment (discussed in this paper) in which the subjects show some degree of cooperation, they manipulate the calibration by adding a correlation of its variables. Despite their efforts, the calibrated model is incompatible with this experiment.

Fehr & Schmidt do not refer to these methodological problems, nor do they lay out an explicit methodology for their project. Fehr & Schmidt do not make it clear what they mean by 'explaining the experiments', they seem to change their methods between papers, but they do not discuss explicitly what they intend to do.

In this paper I critically review the methods used by Fehr & Schmidt in their various articles and appraise the overall explanatory value of their theory. I look mainly at the following 4 papers: [Fehr and Schmidt 1999], [Fehr and Schmidt 2004a], [Fehr, Klein, and Schmidt 2006b] (forthcoming in Econometrica), and [Fehr, Kremhelmer, and Schmidt 2005],

I will show how the absence of methodology lead Fehr & Schmidt to repeatedly make basic logical and fatal methodological errors in their arguments. The result of their efforts is that the calibration and the propositions they prove are of little use, they fail to establish a connection between the theory's predicted equilibria and the experimental behavior. The theory of Inequity Aversion does not provide an explanation of the data.

Fehr & Schmidt aimed to convince their readers that the behavior in many experiments can be explained by their theory. They have failed in their project. The theory of inequity aversion does not further our understanding of the experimental behavior beyond the trivial statement that by manipulating the infinite parameters of the population's preferences it is probably possible to find some compatibility between the theory and any experiment.

In a pamphlet I circulated on the Internet in March 2005 [Shaked 2005], I have shown how the lacunae left by the missing methodology, were filled by rhetoric, hyperboles, overstatement of results and cavalier treatment of data. Fehr & Schmidt have replied to my pamphlet and circulated their response on the internet [Fehr and Schmidt 2005].

It is not easy to separate methodology from rhetoric, both contribute to the real and the perceived explanatory value of a theory. When logical arguments are weak, rhetoric tends to have the upper hand. The reader who wishes to learn about the rhetorical devices applied by Fehr & Schmidt in these papers, can find them in some of the appendices of this paper.

The special nature of this paper requires a large amount of substantiating evidence in the form of citations, quotations and computations. In order to be short and smoothen the reading of these facts, I have put most of the evidence

in extensive appendices. The trusting reader need not read these appendices, although they can be read on their own as an entertaining cautionary tale.

Throughout the paper I refer to the authors as F&S, with apologies to A. Klein and S. Kremhelmer.

2 F&S' Explanatory Methods

2.1 The Theory of Inequity Aversion

F&S propose the following utility function as representing their Inequity Aversion for a population of n individuals:

$$U_i(x_1, x_2, ...x_n) = x_i - \frac{\alpha_i}{n-1}\sum_{j\neq i}\max\{x_j - x_i, 0\} - \frac{\beta_i}{n-1}\sum_{j\neq i}\max\{x_i - x_j, 0\},$$

$$\text{where} \quad 0 < \beta_i < 1, \quad \beta_i \leq \alpha_i$$

For 2 individuals, this becomes:

$$U_1(x_1, x_2) = x_1 - [\alpha_1 \max\{x_2 - x_1, 0\} + \beta_1 \max\{x_1 - x_2, 0\}].$$

The parameter α measures the *envy* of being poorer than another individual, while the parameter β measures the *discomfort* of being better off. The utility function is normalized by the factor $n-1$, where n is the size of the population. The utility function allows an individual to behave altruistically or spitefully depending on the distribution of payoffs.

Each individual is characterized by a pair of parameters (α, β), and the population by a *joint* distribution of α, β.

The theory does not specify the distribution of α, β in the population, it is a theory with, potentially, an infinite number of parameters. By varying the population one can obtain a spectrum of predictions ranging from the competitive to the cooperative. Clearly, for any given game, the theory's prediction will depend on the degree of inequity aversion in the population. If most of the individuals in the population are selfish, the outcome will be close to the competitive selfish equilibrium, while if a large proportion of the individuals is highly inequity averse, the outcome is likely to be cooperative and egalitarian. Thus, for a given measure of cooperativeness it is possible to find a population for which the theory predicts this degree of cooperation.

In their QJE article [Fehr and Schmidt 1999], F&S calibrated the theory by using data on Ultimatum Games, they intended to hold this calibration constant and use the calibrated theory to explain the behavior in a number of experiments (for the little that F&S say about their methodology see Appendix A.1, p. 18).

F&S discuss 4 types of experiments in their paper: markets games with proposers' and with responders' competition, and public good games with and without punishments.

F&S study the following experiments in their article:

1. A Market with Proposers' Competition, [Roth, Prasnikar, Okuno-Fujiwara, and Zamir 1991].

   A number of proposers make offers to a single responder who is restricted to accept (or reject) the *highest offer*. One of the proposers who made the highest offer is chosen at random to divide the surplus with the responder. The outcome in the experiment was the competitive one, the proposers offered all the surplus to the responder, (for the special features of this game, see Appendix A.2, p. 18).

2. A Market with Responders' Competition, [Güth, Marchand, and Rulliere 1997].

   A proposer makes a single offer to a number of responders. Among those who accepted the offer a random responder is chosen to divide the surplus with the proposer. In the experiment about 80% of the responders were willing to accept any offer.

3. Public Good Games without Punishment.

   In these games a number of identical individuals with income y may contribute part of their income towards a public good. A player may contribute an amount g of his income, which becomes an amount ag ($a < 1$) of the public good and is enjoyed by all, including himself In the experiments about 73% did not contribute at all.

4. Public Good Game with Punishment, [Fehr and Gächter 2000].

   This game is like the public good game without punishment, with an additional last stage in which individuals may punish others with a cost to themselves. In this experiment about 80% of the individuals contributed all their income in the first stage.

The Calibration

F&S use data on Ultimatum Game (UG) to calibrate their model. According to the theory, the behavior of an inequity averse responder in the ultimatum game is solely determined by his envy parameter (α) and that of a proposer (who is assumed to know the α distribution of the responders) by his discomfort parameter β. The distribution of β can be calculated from data on the proposers' offers, while the distribution of α can be calculated from data on the responders' acceptance and rejection rates.

All that can be said about a proposer who made an offer $\geq 1/2$, is that his β is greater than 0.5 (Proposition 1, p.826). Unless the experimental data explicitly details how an individual subject behaved as responder *and* as a proposer, the data can provide no information about the *joint* distribution of α, β. The data provides only partial and separate information on each of the marginal distributions of the two parameters.

F&S do not present any individually detailed data, they can therefore compute only the marginal distributions of α and β. In addition, F&S inform us that 40% of the proposers made the offer 1/2, thus the theory, in conjunction with the data, does not pin down the β values of 40% of the proposers, it merely states that $\beta \geq 0.5$.

TABLE III
ASSUMPTIONS ABOUT THE DISTRIBUTION OF PREFERENCES

DISTRIBUTION OF α's AND ASSOCIATED ACCEPTANCE THRESHOLDS OF BUYERS			DISTRIBUTION OF β's AND ASSOCIATED OPTIMAL OFFERS OF SELLERS		
$\alpha = 0$	30 percent	$s' = 0$	$\beta = 0$	30 percent	$s^* = 1/3$
$\alpha = 0.5$	30 percent	$s'(0.5) = 1/4$	$\beta = 0.25$	30 percent	$s^* = 4/9$
$\alpha = 1$	30 percent	$s'(1) = 1/3$	$\beta = 0.6$	40 percent	$s^* = 1/2$
$\alpha = 4$	10 percent	$s'(4) = 4/9$			

Figure 1: **The QJE Calibration** (the marginal distributions) **[Fehr and Schmidt 1999] p.** 844

Clearly, the model is underidentified. The data cannot determine a unique calibration, there is a large set of distributions which are compatible with the data, this leaves F&S the freedom to select one among them.

F&S make no attempt to use statistical, econometric or other scientific methods to determine the distributions of $\alpha, \beta's$ in the population. The values for the distribution of $\alpha's$ cannot be derived from the presented data, and for the 40% of the proposers who made the offer 1/2, F&S simply choose the value $\beta = 0.6$ without justifying their choice.

The final selection of the *two* marginal distributions is presented in figure 1. I will refer to it as the QJE calibration, (for the use of the ultimatum games data for the calibration, see Appendix A.3, p.19).

The value 0.6 for the high $\beta's$ was not chosen arbitrarily. For each game, F&S provide propositions that specify conditions for the existence of an equilibrium which is similar to the experimental behavior in this game. For the experiment of Güth, Marchand, and Rulliere, Proposition 3 (p.832) requires that there should be sufficient individuals with $\beta < \frac{5}{6} = 0.8\bar{3}$. To explain the public good experiment of Fehr and Gächter [Fehr and Gächter 2000] it is necessary that *all* the individuals with $\beta \geq 0.5$ will have $\beta \geq 1 - 0.4 = 0.6$ (Proposition 5, p.841), i.e. there should be no individuals with $0 < \beta < 0.6$. F&S have chosen the value $\beta = 0.6$ which happens to satisfy the two requirements $0.6 \leq \beta < 0.8\bar{3}$. F&S have not used any scientific method to select the value of β which could have assumed any value in $[0.5, 1)$.

Later in their paper, after selecting the calibration, F&S study the public good game with punishment (Fehr and Gächter). To ensure the existence of a cooperative equilibrium in which individuals contribute to the public good (as subjects do in the experiments), there should exist individuals who have *both* high $\alpha's$ *and* high $\beta's$ (the *'conditionally cooperative enforcers'* in Proposition 5). As we have seen, the data of the ultimatum game provides no information on a correlation between the variables. To fit the requirements of the proposition, F&S introduce a perfect correlation between the variables α, β to the calibration of Table III. The new, manipulated, calibration is given by figure 2 (for the way the correlation was introduced, see Appendix A.4, p.19):

%	α	β
30 %	0	0
30 %	0.5	0.25
30 %	1	0.6
10 %	4	0.6

Figure 2: The QJE Calibration with the Added Correlation

Can a calibration that was selected with the help of some data explain this data? When using a calibration obtained from the data on UG to explain the behavior in an experiment A, one should make sure that the data to be explained plays no role in selecting the calibration. If the calibration was selected independently of the experimental data of A, then it may explain the behavior in A. If the selection of the calibration was not independent of A, then it may have been selected for the purpose of fitting the predictions of the calibrated theory to the data of A. In that case, the calibration cannot explain the behavior in A. The most that can be claimed is that a distribution has been found which is compatible with both data sets of *UG and* of A. Clearly the first case (explanation) is logically stronger, it implies the second (compatibility).

The selection of a calibration can be said to be independent of an experiment A, if it was made before the data for A was available, or alternatively, if a well established statistical or econometric method has been used to select the calibration. In all other cases the calibration cannot be assumed to be independent of the data and it would be wrong to claim that it can explain the data that was instrumental in its selection.

The data of the experiments was available to F&S when they selected the calibration. Under these circumstances it cannot be said that they have chosen this value independently of the data.

Although there is no reference to it in the paper, the choice of $\beta = 0.6$ was not a mere coincidence. In their reply to my pamphlet [Fehr and Schmidt 2005], F&S admit that their choice of the value $\beta = 0.6$ for the calibration was influenced by the data of the experiments that were claimed to be explained by this same calibration They justify their choice by explaining that $\beta = 0.8$ is an extreme β value, that $\beta = 0.6$ seems a plausible value, that it is more realistic to assume a continuous distribution of $\beta's$ and that it seems plausible that someone with a high β has also a high α (p. $5, 6, 7$ in [Fehr and Schmidt 2005]). All these arguments come too late in the process, F&S should have introduced them as additional assumptions to their theory, before considering the particular experiments (for F&S' version of how they selected the calibration, see Appendix A.5, p.19).

The data of two experiments (Güth, Marchand, and Rulliere and of Fehr and Gächter) was instrumental in selecting the QJE calibration. Therefore, the calibrated theory cannot explain these experiments. The most that can be said is that F&S have found a calibration which is compatible with the data of these experiments.

The Public Good Experiments.

Despite these stratagems, the calibrated theory is not compatible with any of the public good experiments. For Public Good Games without Punishments, the calibrated theory predicts that nearly 100% of the individuals should not contribute to the public good. However, in the experiments, only about 73% did not contribute to the public good (Table II, p.838). The discrepancy between the theory and the data is very large, $\frac{100-73}{73} = 36.9\%$, the theory is not compatible with the data, (for F&S' reference to this gap, see Appendix A.6, p.20).

For the Public Good Game with Punishment [Fehr and Gächter 2000], Proposition 5 requires that the population consists of **two types** only, the *'conditionally cooperative enforces'* - individuals (with $\beta \geq 0.6$ and a correspondingly high α), and selfish ones (with $\alpha = \beta = 0$). The QJE calibration has 4 types, about 30% of the population have intermediate values of α, β and are are neither selfish nor *'conditionally cooperative enforcers'*. Proposition 5 does not apply to the QJE calibration, it is based on a different calibration, and as a result it has not been shown that the calibrated theory is compatible with the behavior in this experiment.

A Summary of the Explanation Provided by the QJE Article.

To summarize the achievements of the QJE article, we go through the list of experiments that were discussed in that article:

1. F&S succeeded in explaining the behavior in the market experiment with proposers' competition [Roth, Prasnikar, Okuno-Fujiwara, and Zamir 1991]. F&S prove that *irrespective* of the composition of the population, the predicted outcome is always the competitive one. However, the rules of this game do not allow the players to freely practice their inequity aversion if they have any (see Appendix A.2, p.18).

2. The data of the market experiment with responders' competition [Güth, Marchand, and Rulliere 1997] influenced the calibration selection, therefore, it cannot explain the data.. However, the data is compatible with theory's prediction.

3. The behavior in the public good experiments without punishment has not been shown to be compatible with the theory's prediction.

4. The data of the public good experiment with punishments [Fehr and Gächter 2000] was instrumental in selecting the calibration. It was used to determine the parameter values and later their correlation. The proposition, which should have provided a link between the calibrated theory and the experiment, does not apply to the calibration. F&S failed to show that the data is even compatible with the calibration. Note that this is the only experiment, among the four discussed in this paper, in which the players cooperate.

The calibrated theory succeeded in explaining one competitive situation and be compatible with the data of another competitive game, none of the public good games has been shown to be compatible with the calibrated theory. F&S were particularly interested in providing a unified theory which will simultaneously explain the free riding in public good games without punishment and the

contributions in the public good game with punishment. Their theory failed to explain or even be compatible with these experiments. F&S failed to produce a single population that can explain the experiments they considered in [Fehr and Schmidt 1999]. The theory did not add anything to our understanding of the experimental behavior. It has little or no explanatory value.

2.2 Explaining Contract Choices.

Since introducing the theory of inequity aversion in [Fehr and Schmidt 1999], F&S conducted a number of experiments concerning choice between various contracts, these were analyzed and explained with the calibrated inequity aversion theory.

F&S have publicly declared (in [Fehr and Schmidt 2005]) that it is on the basis of these experiments that they felt confident to claim that their calibrated QJE model yields *quantitatively accurate predictions*, (for F&S' declarations, see Appendix B.1, p.20).

In this section I look at the methods applied by F&S in 3 papers: [Fehr and Schmidt 2004a], [Fehr, Klein, and Schmidt 2006b] (fothcoming in Econometrica), and [Fehr, Kremhelmer, and Schmidt 2005] I refer to these papers as the Contract Papers.

The 3 contract papers have the traditional structure of an experiment followed by a theory, and F&S claim to explain and interpret the experimental results with the calibrated model (for F&S' claims, see Appendix B.2, p.21).

A number of points should be noted about these papers:

1. Despite their claim that these papers support the QJE calibration, none of the papers uses the QJE calibration, F&S have switched to use a different calibration.

2. Although the calibrated theory makes some clear and simple predictions about the fundamental behavior in equilibrium, F&S do not test these predictions. We test these predictions and find that they do not match the data.

3. F&S select some general patterns (*'qualitative patterns'*) from the data and the theory. They claim that these patterns agree in the theory and the data, and on the basis of these patterns they conclude that the data is 'largely consistent' with the theory. We study these patterns and find that none of them shows that inequity aversion is relevant to the experiments.

The models of these three papers differ in their details but the reasoning and intuition driving them is the same. I concentrate here on that part of the experiments in which players choose between Bonus and Incentive contracts [Fehr, Klein, and Schmidt 2006b], between A-ownership and Joint ownership [Fehr, Kremhelmer, and Schmidt 2005], or between a Piece-wise and a Bonus contract [Fehr and Schmidt 2004a]. I use the terms and language of [Fehr, Klein, and Schmidt 2006b], but, with slight variations, these apply to the models of the other two papers.

A principal offers a contract to an agent. The agent can then exert a costly effort which produces a payoff for the principal. The principal may offer the agent an Incentive, Trust or a Bonus contract. In an incentive contract the principal may invest in verification technology, he names a wage, he demands an effort level and specifies a fine to be paid if the agent made a lower effort. In a Bonus contract the principal names a wage, an effort level and a bonus which he may pay. Here, neither the agent's effort nor the principal's bonus are contractually enforceable. A Trust contract is like the bonus contract without the last stage, the stage in which the principal may pay a bonus.

The papers contain some other experimental treatments, but I will refer only to the Bonus/Incentive or Bonus/Trust treatment, and in particular I will look at the behavior under the bonus contract. The Bonus contract game is a natural environment to test the theory. In all other contracts (Incentive and Trust contracts) F&S tie the hands of the players and prevent them from making a unilateral payment to others at the end of the game. This can be easily achieved in an experiment, but if the preferences of the inequity averse individuals are to be taken seriously, this seems a rather unnatural restriction. It is not easy to imagine a situation in which a player is prevented from anonymously mailing some money to another, if he really wishes to do so. If this payment is permitted, then an inequity averse principal will obey his nature and compensate his agent whenever he (the principal) has a higher payoff. If all agents are aware that (some) principals will pay a bonus, the situation is equivalent to a bonus contract.

We now investigate the methodology applied by F&S in these papers and evaluate its explanatory value.

Tailoring Calibrations to Experiments.

The QJE calibration is not used in the contract papers. The population, is assumed to consist of 60% selfish individuals (with $\alpha = \beta = 0$), and 40% highly inequity averse, fair individuals, with high but unspecified inequity aversion parameters $\alpha, \beta > 0.5$. To create this new set of distributions F&S have eliminated 30% of the QJE population, (those individuals with intermediate values $\alpha = 0.5, \beta = 0.25$) and correspondingly increased the weight of the selfish individuals. They also allow the inequity averse individuals to have any (high) values of α, β and not only $\beta = 0.6$ and $\alpha = 1$, or $\alpha = 4$ as in the QJE distribution. Clearly, there are many types of fair players, depending on their α, β, but F&S ensure that all of them behave identically in equilibrium. I refer to this population as a $40 - 60$ distribution (for some properties of the $40 - 60$ distribution and how F&S change the calibrations in their appendices, see Appendix B.3, p.21).

Obviously, the QJE calibration is *not* one of the $40 - 60$ distributions. In the QJE calibration 30% of the population has intermediate values of both α and β, which are neither 0 nor higher than 0.5.

The $40 - 60$ distribution with its unspecified values of α, β is incompatible with 3 of the experiments in the QJE paper [Fehr and Schmidt 1999], (see Appendix B.4, p.22).

There is no attempt to show that the QJE calibration is compatible with the data of the contract experiments. The move from the QJE calibration to the $40 - 60$ distribution can be justified if the individuals with intermediate

inequity aversion (who no longer exist in the new distribution) would behave, in the contract games, like the selfish individuals with whom they were grouped. But this is not the case, it is easy to verify that in the proposed equilibrium (given the equilibrium behavior of the *principals*), the *agents* with the QJE intermediate values of $\alpha = 0.5$ (and $\beta = 0.25$) will behave like the fair agents (with $\alpha > 0.5$) and not like the selfish ones (with $\alpha = 0$).

F&S make no attempt to reconcile the new $40 - 60$ distribution with the experiments discussed in [Fehr and Schmidt 1999]. F&S do not discuss the problems of switching between calibrations, and when they present the new distribution, they downplay the differences between it and the QJE calibration by using ambiguous and ill-defined terms (for F&S' description of the new calibration, see Appendix B.5, p.22).

In their attempts to explain experimental behavior, F&S fitted different calibrations to different experiments. The experiments in [Fehr and Schmidt 1999] were 'explained' with the QJE calibration, while the contract experiments use the new distribution.

Is there any explanatory value in fitting different populations to different experiments? Tailoring populations to experiments has no explanatory value since it is obvious that by selecting a suitable population the theory's prediction can freely change.

Moreover, explaining different experiments with different populations amounts to saying that the degree of inequity aversion in a population changes with the situation (game) it faces. While this may be true, this statement does not add anything to our understanding of the experimental behavior, unless accompanied by a theory predicting how the degree of inequity aversion changes with the situation. Such a theory will have to explain how facing a particular game causes individuals to change their preferences, it will also have to explain how this changing distribution becomes known to all players, (F&S assume that, in equilibrium, all players are familiar with the correct distribution of inequity aversion in the population). In effect, such a theory will select an equilibrium according to a social norm. F&S have not yet proposed such a dynamic theory of inequity aversion.

The Neglected Predictions and the Qualitative Patterns.

We come now to the 2^{nd} and 3^{rd} points we noted earlier, that F&S test only selected 'qualitative' predictions of their theory. In the theoretical parts of these papers, F&S provide minutely detailed propositions which make very precise predictions of the fundamental behavior in the studied games. The propositions predict the proportions of agents who exert low effort and of principals who reward cooperation, these fundamental properties determine the players' payoffs and the contract they choose. In addition, these predictions are simple, straightforward and can be easily tested, yet, F&S do not test them. All they test are some general behavioral patterns, which they term *'qualitative'*, these include the choice of contracts in the theory and the experiment. On the basis of the 'qualitative' features F&S claim that the theory is *'largely consistent'* with the data. Below, we describe the theory's predictions and the qualitative patterns, we test them, and find that the theory's predictions are incompatible with the data and that those qualitative patterns that agree with the theory do not add to the theory's explanatory value.

Why do F&S test the qualitative patterns, but stop short of testing the theory's quantitative predictions? Why are some of the theory's predictions valid for testing and others are ignored? Is there a logical distinction between the two which justifies the different treatment? F&S do not refer to these questions and they offer no answers, it is left to us to find the logic in their scheme.

Indeed, the theory's predictions are directly based on the $40-60$ calibration which F&S use for these games. F&S went out of their way to convince us that this calibration is relevant to their model, but when it comes to the test they treat it as irrelevant, (for F&S' statements about the relevance of the $40-60$ distribution, see Appendix B.6, p.23).

In two of the contract papers F&S inform us, rather casually, that the *qualitative results* of their propositions hold for a wider class of calibrations, (for a description of these calibrations and how F&S present them, see Appendix B.7, p.23). A careful examination of the papers reveals that the proofs of the proposition do not mention other calibrations, nor do F&S make any use of this wider class in their arguments.

What is the purpose of this information? If the $40-60$ calibration explains the data, why are other calibrations required? Are we being told, in a subtle, indirect way, that the $40-60$ calibration may not be compatible with the data, but that some other calibration may be more successful? Whatever this information is meant to convey, F&S do not make any use of it, and they do not test the more detailed predictions of any of the calibrations.

In the equilibria considered by F&S, all individuals of a certain type behave identically. Thus there are *at most* 4 cases, a fair principal meets a fair agent, a fair principal meets a selfish agent, etc. The equilibrium specifies the frequencies of these cases in the population, the data can be easily tested for these frequencies. In the equilibrium, all principals offer the bonus contract, all selfish agents expend high effort and all fair agents a low one, selfish principals pay no bonus and fair principals reward only selfish agents.

The intuition for this equilibrium is straightforward and is clearly presented in all the papers. An agent who makes a high effort reaches an unequal allocation in which he has the short end. Under a bonus contract, the agent is sure to be compensated by the fair principals (those with a built in preference for egalitarian allocations).

Since not all of the principals are fair, the agent's expected allocation is not completely egalitarian but close to it. This partial bonus may suffice to induce a selfish agent to expend high effort, whereas inequity averse agents may shun this expected unequal allocation and opt for a more egalitarian allocation with a lower, inefficient, effort level.

F&S compute their equilibrium only for a particular type of calibration. They do not consider a population with more than two types (the selfish and the fair), nor do they indicate how their equilibrium could be generalized to populations with more varied types. In this model, a population with multiple types is more than a theoretical nicety. The experimental data suggests that there is no perfect correlation between the α, β parameters and that there are more than two types in the population. Moreover, the equilibrium is not robust to minor changes in the games' parameters. The parameters of the games, the production and cost functions, were carefully selected to support the equilibrium

by making all fair players choose the low effort. Minor changes in the parameters can destroy this equilibrium by inducing some of the fair agents to cooperate. To describe an equilibrium with cooperation for such games would require a calibration with more than two types (for a detailed description of how fragile the equilibrium is, see Appendix B.8, p.24).

In [Fehr, Klein, and Schmidt 2006b] F&S admit that their individual subjects do not behave according to inequity aversion theory, but they claim that the aggregated data matches the theory. In appendix B.9, p.25, we test the data and find that even on an aggregated level the data does not match the theory. We show that the fractions of fair agents (those who expend low effort), and of fair principals (those who reward high effort) do not match the calibrated theory's predictions. We find that the data suggests that the population in the experiments is incompatible with the $40-60$ calibration. There is no perfect correlation between fair agents and fair principals, and the percentages of fair agents and fair principals, as suggested by the data are far removed from those of the calibration. Since in some experiment the percentage of fair agents is much higher than that of fair principals while in another the reverse holds, the data seems to suggest that no single distribution can explain all experiments.

In the experiment of [Fehr, Klein, and Schmidt 2006b], the fraction of fair principals, suggested by the data, is so low that the theory, calibrated with a population with that fraction of fair principals, predicts that under a bonus contract there will be very little cooperation (selfish agents should exert rather low effort, much lower than in the actual experiment). We also show that bonuses are not paid to equalize payoffs, as the theory assumes, that agents payoff do not match the theory and that the proportion of bonus to total compensation is not as the theory predicts.

The tests show that the data does not match the fundamental behavior predicted by the theory. The factors that determine the attractiveness of a bonus contract relative to another contract are, among others, the fraction of principals who pay a bonus, the magnitude of the bonuses and the degree of agents cooperation. All these factors are incompatible with the experimental behavior. The theory, its propositions and the choice of contract it predicts are all irrelevant to the data.

Let us now consider in detail the 'qualitative patterns' which according to F&S show that the theory is 'largely consistent' with the data. The detailed tests and the relevant computations can be found in appendix B.9, p.25.

The main qualitative feature is that the principals choose the same contracts in the theory and in the experiments: Bonus contract is preferred to Incentive contract and Incentive contract to Trust contract. F&S commit a grave logical error: the final conclusion of the theory, the choice of a contract, depends on the underlying fundamental behavior under the two (or more) contracts, and this behavior does not match in the theory and the data. If the final choice happens to be the same in the theory and the experiment, it is not because of inequity aversion behavior. If inequity aversion were the underlying reason for the experimental contract choice, then the fundamental patterns of behavior in the experiment would have followed the theory.

It is not only the subjects' *individual behavior* that does not fit the theory, as F&S themselves admit, it is also the fundamental *aggregate behavior* that does not match the theory (percentages of cooperators, bonuses and payoffs). In what

sense, then, does the theory *rationalize the quantitative facts* ([Fehr, Klein, and Schmidt 2006b], p.25)? Can one conclude, despite all these discrepancies, that it is inequity aversion that causes the subjects to choose the way they have?

Imagine that we have a theory about the moons of the planet Jupiter. The theory states that Jupiter's location in the solar system caused it to have high temperature, and that this, in turn, caused explosions which created its moons. We now observe the planet, we confirm that it has moons, but we find that Jupiter's temperature was never high, and we find no evidence of violent explosions. On the basis of the confirmation of the *qualitative* pattern (that Jupiter *has* moons) we conclude that the theory is *largely consistent* with the observations. As in this somewhat simplified example, inequity aversion theory offers no explanation of the data.

The population in the experiments is incompatible with the calibration assumed in the propositions and used to predict the equilibrium behavior. The propositions and the theory's predictions do not apply to the population of the experiments.

For example: In the experiment of [Fehr, Klein, and Schmidt 2006b], the population in the experiment has few fair principals (who pay a reward). There are so few fair principals, that according to the theory, there should be very little cooperation in the bonus contract (the effort levels chosen by all agents will be rather low $e \leq 3$.) If this is the case, it is very likely that the theory predicts that, between a bonus and an incentive contract, the principals will choose an incentive contract, contrary to what subjects do in the experiment. By ignoring the relevant tests, F&S have applied an irrelevant theory to their data.

A second qualitative feature considered by F&S is that in the experiments the paid bonuses increase with the expended effort, and that the bonus forms a substantial part of the agents' payoff.

The only theoretical reason for an inequity averse principal to pay a bonus is to equalize his payoff to that of his agent. For a given wage rate, this implies that the bonus increases with effort *in a particular way*, and changes with the level of effort (the derivative of the bonus w.r.t. effort is determined by the production and cost functions). Note that the property of equating payoffs is independent of the calibration.

Rather than test the fundamental assumption of the model that bonuses are paid to equalize payoffs, F&S test only whether bonuses increase with the effort expended by the agent. They run a regression and find that the *average* increase of bonus per unit of effort is such that a rational selfish agent would indeed choose the equilibrium level of effort. The test should have been whether the bonus varies with effort as the theory predicts.

F&S commit the same logical error as in the previous qualitative pattern. Even if it is found that bonuses increase with effort, this does not mean that the principals follow the inequity aversion theory. The real test should be whether bonuses attempt to equalize the payoffs of the principal and his agent. For two of the experiments, [Fehr and Schmidt 2004a], [Fehr, Klein, and Schmidt 2006b], we show (see appendix B.9) that the bonuses paid in the experiment do not equalize the payoffs of the principal and the agent. If, despite that, bonuses increase with effort, it does not demonstrate that it is because individuals are inequity averse. (It is also trivially true that if for some experiment *all* positive

bonuses equalize payoffs, one can easily find a population with an appropriate proportion of selfish to fair principals which will be compatible with this data).

The same holds for the statement about the share of the bonus in the agents' payoffs. The underlying reason for paying the bonus is different in the theory and the data, hence the observation that the bonus is a substantial part of the payoff does not show that players are inequity averse. It is therefore not surprising that a test and comparison of the theoretical and experimental share of bonus/payoff finds them to be very different and incompatible.

Another qualitative pattern noted by F&S is that, in the experiments, bonus contracts have outcomes which are more efficient than those of a trust contract. Although all the theoretical results in the papers depend strongly on F&S' choice of the game parameters (See Appendix B.8, p.24), F&S present one single proposition which is robust to changes in the game parameters. They show that a bonus contract is (weakly) more efficient than a trust contract. The proof holds for any parameters of the game (production and cost functions) and any population with $q \in [0,1]$ fair individuals ($\alpha, \beta > 0.5$) and $1-q$ selfish ones. (see proposition 6, p.$A-8$ of the appendix [Fehr, Klein, and Schmidt 2006a]).

In a trust contract the principal announces a wage and the agent reacts by choosing an effort level. In a bonus contract there is an additional move by the principal, in which he can pay the agent a bonus. In a trust contract the inequity averse principal, against his preferences (his β parameter), is not allowed to pay the agent and is prevented from reaching an egalitarian allocation. It is therefore clear that for any wage paid in the first stage and any belief q of the agent, the agent will be (weakly) more cooperative and efficient under the bonus contract.

Stripped of its contractual jargon, this statement merely says that the bonus game has an additional stage in which the mover (if he is better off than the other) may practically announce the two players to be a team and share any profits they may have. Whatever the actions taken by the principal in the first stage, the extra move added to the bonus contract cannot reduce efficiency, and it certainly increases efficiency when q is sufficiently large (F&S provide a scholarly, 3 pages long, proof of this observation).

This qualitative feature (that bonus contracts are more efficient than trust contracts) is really about the agents' *belief* about the existence of individuals who may make use of the additional move. It is *not* about the existence of such individuals. If, in some experiment, we observe that bonus contracts are more efficient, it shows that, like F&S, some agents believe that some of the principals will pay a bonus. It does not confirm that individuals are inequity averse, only that some individuals believe that others may be. F&S have not tested the agents' beliefs, and there is no reason to believe that the agents' beliefs are correct (the experimental behavior does not match the predicted equilibrium).

Yet another qualitative pattern presented by F&S is that although the bonuses paid in the experiments are high, some agents do not cooperate and expend low effort. According to the theory, those individual who choose a low level of effort, do so because of their high α value. The observation that some agents do not cooperate does not confirm the theory, the agents may shirk because they do not believe that there are any inequity averse individuals in the population. As the previous qualitative pattern this one is about agents' beliefs about the existence of principals who pay the reward.

Although F&S avoid any quantitative tests, in [Fehr, Klein, and Schmidt 2006b] they quantitatively compare the data and the theory. They discover that the average wage offered in a bonus contract, the average bonus and the average effort level agree in the experiment and the theory (the averages are based on the $40-60$ distribution).

But averages are not a satisfactory measure, "they may frequently hide differences at a more disaggregated level", (a quotation from [Fehr and Schmidt 2004a], p. 463). We have shown that the underlying empirical behavior does not agree with the theoretical predictions. The wage paid and the effort expended depend on the fractions of fair principals and fair agents and these are different in the theory and the data. The rationale for paying bonuses is not the same in the theory and practice. In addition, F&S admit in the paper that the subjects' individual behavior does not match the theory (p.25), it is therefore clear that the accord between the averages that was noted by F&S is a mere coincidence. Based on this accidental equality of averages, F&S exclaim 4 (!) times in their paper that their theory *provides surprisingly accurate and remarkably precise quantitative predictions of the details of the bonus contract* ([Fehr, Klein, and Schmidt 2006b], pps. $4, 21, 25, 29$) These statements are clearly misleading.

The following test would be a more suitable one: according to the theory (with the $40-60$ distribution), the average bonus paid to an agent **who made a high effort** is $0.4 * 25 = 10$, the empirical average bonus paid to an agent who made a high effort ≥ 5 is $\frac{1811}{127} = 14.26$, (Table V, [Fehr, Klein, and Schmidt 2006b], or $\frac{2767}{213} = 13$ according to Table 5 [Fehr, Klein, and Schmidt 2006a]). This is a discrepancy of the order of $30-42\%$.

In all their references to these accurate predictions F&S describe them as 'surprising'. What is it that F&S find so surprising about these averages? The above quotation from their previous paper shows that F&S are aware that two completely different distributions may share the same average. Could it be that they use 'surprising' as a subtle reminder to the reader of the serendipitous and accidental nature of this result?

F&S are apparently so confident that it suffices to confirm the compatibility of some 'qualitative patterns', that they do not even propose to test the fundamental predictions of the theory. Obviously, the referees and editors of the journals that publish these papers, share their confidence, since they did not require F&S to run these tests. As a result, none of the qualitative patterns shows that inequity aversion is relevant to the data. Figure 3 describes the logical structure of the predictions, and the mismatch between the theory and the data.

A Summary of the Explanation provided by the Contract Papers.

F&S abandoned the QJE calibration and started using a new one. In effect, they tailor different calibrations to different experiments. They claim that the theory predicts some general features of the experimental data but they avoid testing the theory's simple and fundamental predictions, which turn out to be incompatible with the data. The calibrated theory predicts behavior that is so far removed from the experimental data, that it is clear that the experimental population and the calibration are very different. This means that the theory's predictions do not apply to the population of the experiments.

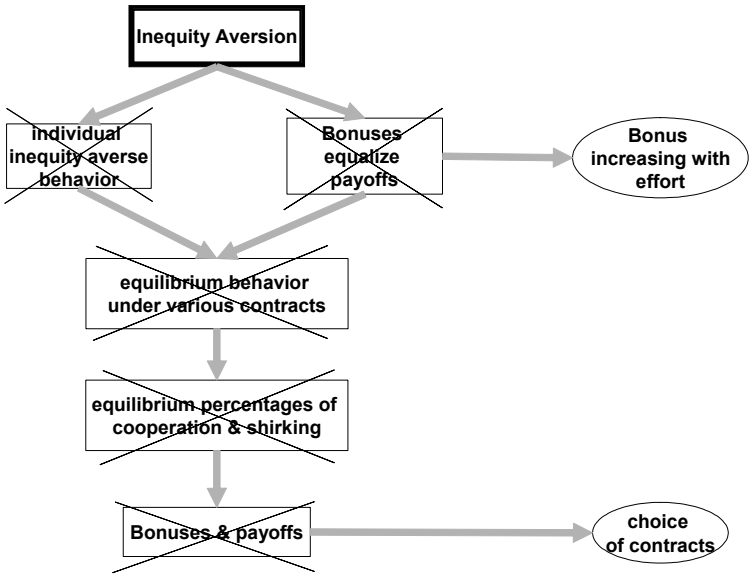

Figure 3: The crossed boxes contain theoretical predictions that do not match the data. Those in the oval boxes agree with the data

Moreover, the data suggests that there is no single calibration that is compatible with all the experiments, each experiment requires a completely different calibration. The qualitative patterns considered by F&S do not show that the experimental behavior is due to inequity aversion preferences. The theory provides no explanation of the experimental data, it has no explanatory value.

In their latest paper [Fehr, Klein, and Schmidt 2006b], F&S conclude by admitting that their model fails to provide a good description of the behavior of individual subjects, but that 'on average subjects behave *as if* they were motivated by inequity aversion'. We have seen that it is not only the individual behavior, but also the aggregate behavior that is incompatible with the theory. One would be justified in using the *as if* approach if the final conclusions of the theory agree with the data, but currently there are no available methods to test the theory's intermediate predictions. However, in this case, the theory's intermediate predictions can be directly tested and are found to be completely incompatible with the data. It is meaningless and misleading to claim that the subjects behave *as if* motivated by inequity aversion and that the theory *'helps to organize and interpret the data'.*

3 Conclusion

F&S wanted to present a theory that provides a unified explanation to many different situations which cannot be explained by the traditional selfish preferences assumption applied to one-shot games. However, F&S have not laid out a consistent methodology for their project. The absence of a sound foundation caused them to commit grave logical errors, and as a result their project has failed. While clear methodology is absent, the reader is confronted with am-

biguous phrases and misleading statements not only in their papers but also in the public and open discussion of their theory (for a description of the rhetoric in F&S papers, see Appendix B.10, p.31).

The sum total of F&S work amounts to manipulating calibrations in an attempt to fit data, tailoring distributions to experiments, ignoring predictions which do not fit the data and relying on general 'qualitative patterns' which do not support the theory. By using these questionable methods, F&S have failed to show that the calibrated theory of inequity aversion is relevant to the experiments.

The general treatment of the calibration in these papers, may indicate that the calibration was not meant to be taken too seriously. Perhaps the aim of these papers was merely to state that experiments show that some individuals are motivated by inequity aversion. But if neither the number of these individuals nor their degree of inequity aversion is specified, then this statement is trivially true. By fine-tuning the infinite parameters of the model nearly anything can be explained. There is no need to run any experiments nor prove any theorems to understand this rather simple point.

Otherwise, the theory has added nothing to our understanding of the experiments, F&S have not advanced us beyond the obvious observation that the theory can make a wide range of predictions depending on how inequity averse the population is.

It seems that the only purpose of the theory in these papers is to provide a pretense of scientific veneer to the authors' conjectures about the reasons for the subjects' behavior in the experiments. However, the experiments and F&S' conjectures are interesting, intriguing, and do not require the superfluous theory. The experiments and F&S' conjectures could stand on their own and possibly be published in a journal that puts less emphasis on needless theory, provided they are presented for what they really are: an erudite speculation.

I have no doubt that F&S and the hundreds of their followers firmly believe that it is the existence of inequity averse individuals that is responsible for the experimental behavior in many games, and indeed they may be right. But by using dubious methods, F&S have failed to convince the uninitiated.

F&S devoted an impressive number of pages to prove a large number of propositions applying the theory of Inequity Aversion to various games. While these propositions may be of some theoretical interest, they are superfluous since the equilibria they describe do not apply to the data. F&S would have done better proving less theorems and devoting some serious thought to the methodology of their project.

Appendix

A The QJE Paper

A.1 The Methodology.

In [Fehr and Schmidt 1999], F&S do not inform us what their methodology is. It is only in a later survey article [Fehr and Schmidt 2003] that we learn that F&S intended to keep the calibration in their QJE paper constant:

> *"Using the data that is 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 2003], p.222.

F&S have not changed their minds about what they did in their QJE article. In a recent survey article, soon to be published in the Handbook on Reciprocity, Gift-Giving and Altruism [Fehr and Schmidt 2006], F&S repeat these statements nearly word for word, referring to their QJE article, they say:

> *"Fehr and Schmidt choose a distribution for α and β that is consistent with the experimental evidence of the ultimatum game.* ***Keeping this distribution fixed****, they show that their model yields surprisingly accurate predictions across many bargaining, market and social dilemma games"* [Fehr and Schmidt 2006], p.26.

The only noticeable difference between the two quotations is that the *'quantitatively accurate predictions'* of the World Congress in 2000, turned into: *'surprisingly accurate predictions'* in the year 2006, (for other artful uses of 'surprise', see p.15).

A.2 A Market with Proposers' Competition.

It is worth noticing that in all their papers on inequity aversion, this experiment is the only one that is *fully* explained by the theory. Its data was not instrumental in selecting the calibration, and the theory predicts the experimental data perfectly. Moreover, the theoretical prediction is independent of the calibration.

The reason F&S included competitive games in their analysis was to demonstrate that their theory does not clash with the established experimental results on competitive games. In competitive situations subjects, usually, play competitively. This experiment can, of course, be explained by the traditional selfish assumptions, but F&S wanted to show that in such a competitive situation even highly inequity averse individuals play competitively. To demonstrate their point they should have chosen an environment in which the egalitarian preferences of the players can be fully expressed and are not suppressed by the rules of the game.

Recall that the rules of this game strip the responder of his inequity aversion features. He is restricted to consider only the *highest* offer, but a person with

egalitarian preferences (high β) may wish to choose a more egalitarian lower offer. If the game is changed so as to allow the responder to freely choose one of the offers, then it will no longer be possible to obtain a competitive outcome with no restrictions on the calibration. Analogous to the market game with responders' competition, here, in order to obtain a competitive outcome, the responder will have to be sufficiently selfish.

A.3 Selecting the Calibration: The Ultimatum Game Data.

F&S provide only partial data about proposers' offers in the UG, and they provide no data on the responders actions. What little data they present is insufficient to establish their calibration.

The data on proposers' offers (Table I, p. 827 [Fehr and Schmidt 1999]) is incomplete. It does not include intermediate offers (between 0.2, and 0.4) which were made by 25% of the proposers, nor can we learn from it that 40% of the proposers offered an equal split. The information about offers in the interval $(0.4, 0.5)$ is aggregated and amounts to 70% of the offers.

As a source for data on the responders actions, F&S refer to [Roth 1995], but they do not tell us where in this 95 page long article the data is to be found. The data which F&S need for their purpose cannot be found in the paper, some of it may be extracted from the various diagrams in the paper.

The distribution of α was supposed to be derived from the data on Ultimatum game, but when they present the distribution, F&S use the following wording, which suggests that the choice of the values for α was rather unsystematic:

> *"Thus, we* ***(conservatively)*** *assume that* 10% *of the subjects have* $\alpha = 4$ "
>
> *"Another, typically much larger fraction of the population insists on getting at least one-third of the surplus, which implies a value of* α *which is equal to one. These are* ***at least*** 30 *percent of the population."*
>
> *"Another,* ***say,*** 30 *percent of the subjects insist on getting at least one quarter, which implies* $\alpha = 0.5$ *"* (pp. $843 - 844$).

A.4 The Correlation.

The correlation between α, β, which is essential for explaining the Fehr and Gächter experiment, was introduced in the appendix (p.864, [Fehr and Schmidt 1999]). There is no mention of it in the main text of the paper. Those readers who did not read the appendix are not even aware that a correlation has been added to the calibration of Table III. In fact, the reader is mislead to believe that the calibration has not been changed, since F&S tell him, when discussing the Fehr and Gächter experiment (on p.846), that it is consistent with the calibration of Table III (which has no correlation).

A.5 Selecting the Calibration: The Use of the Data to be Explained.

In their reply to my pamphlet ([Fehr and Schmidt 2005], p.7), F&S refer to the choice of $\beta = 0.6.$ for the calibration. They admit that they have chosen this value to be consistent with the required value of proposition 5 :

> *"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."*

It seems that the calibration was selected by using the data of the Fehr and Gächter experiment, but there is no mention of this in the paper.

The choice of $\beta = 0.6$ also satisfies the condition $\beta < 0.8\overline{3} = 5/6$. For the theory to be consistent with the data of [Güth, Marchand, and Rulliere 1997], there should not be too many individuals with $\beta \geq \frac{5}{6} = 0.8\overline{3}$. Although this condition is explicit in Proposition 3 $\left(\beta < \frac{n-1}{n} = \frac{5}{6}\right)$, F&S do not refer to it in their calculations. They fail to mention that it is their choice of $\beta = 0.6 < \frac{5}{6}$ which ensures that *all* individuals satisfy this condition. In their reply to my pamphlet (p.6) F&S justify their choice by claiming that values of $\beta \geq 0.8$ imply *'an extreme degree of inequity aversion'*. But their theory allows β to assume any value < 1, the notion of unacceptable 'extreme degrees of inequity aversion' is not mentioned in any of their papers, and must have been introduced specifically for this argument.

A.6 The Treatment of Data.

In Public Good Games without Punishment, the gap between those who did not contribute anything in the experiment (73%) and the theory's prediction (100%) is of the order of magnitude 36%. Despite this discrepancy, F&S declare that *'it seems fair to say that our model is consistent with the bulk of individual choices in this game'* (p.845).

They refer to this gap in a footnote (footnote 21, p.845), where they make two points about it. The first is that the gap can be partly closed by theories of fundamental randomness of human choice, but they leave this research to the future. The second point is that the gap can be reduced by a significant fraction of the players who made very small contributions to the public good. However, they do not produce the data, nor do they provide any information about the size of this significant fraction of players. This seems particularly odd since the information could have been added in one single column to Table II (p.838), whose sole purpose is to summarize the data of these games.

The absence of this information is even more puzzling since E. Fehr refers to this fraction of subjects in other papers. In two of his papers, E. Fehr refers to Table II as a meta-study of public good games. He makes specific references to the significant numbers of individuals who made small contributions and claims that this was found by the authors (F&S) of this meta-study. He does not tell his readers that this data is not available in the meta-study and that it only makes a fleeting appearance in a footnote, [Gintis, Bowles, Boyd, and Fehr 2003], p.160, [Fehr and Gächter 2000], p.983.

B The Contract Experiments

B.1 Quantitatively Accurate Predictions.

In their QJE paper [Fehr and Schmidt 1999], F&S describe their calibration and their explanation process as *'crude' 'rough'* and as a *'first test'.* Later, in their invited survey paper to the 8th World Congress of the Econometric Society, 2000 ([Fehr and Schmidt 2003] p. 222), when they describe their own work, F&S claim that *"Keeping this distribution* (the QJE calibration) *constant, they* (F&S) *show that their model yields* ***quantitatively accurate predictions*** *across many bargaining, market and co-operation games".*

In my pamphlet, I challenged F&S to explain how their 'crude and rough tests' could have produced *'quantitatively accurate predictions'.*

F&S promptly replied to the pamphlet and listed 3 experiments on contract choice which were not available when they wrote their QJE paper, but were discussed in their invited paper to the World Congress ([Fehr and Schmidt 2005] p.8). They state that their results support the model of inequity aversion as well as the calibration that they used in QJE.

It is these 3 experiments that gave F&S the *"confidence to claim that our model 'yields quantitatively accurate predictions across many bargaining, market, and co-operation games'."*

In their reply, F&S list two more papers which allegedly support their theory and which were not discussed in the invited paper: [Fischbacher, Fong, and Fehr 2003], [Fehr and Schmidt 2004b]. I do not include these papers in my discussion here, since the first applies quantal response equilibria, and it is beyond the scope of this paper to consider the explanatory value of quantal response theory, and the second is not relevant to the calibrated theory.

Note that in their reply, F&S refer to older versions of the 3 contract papers, but there is no essential difference between the old and new versions as far as the experiments and the calibrations are concerned.

B.2 Inequity Aversion Theory Explains the Contract Experiments.

Although F&S do not test the theory's predictions they claim that the theory explains the data.

In two of the contract papers F&S state that they intend to explain and interpret the experimental results with their theory. ([Fehr and Schmidt 2004a] p.456, [Fehr, Kremhelmer, and Schmidt 2005], p.2).

In [Fehr, Klein, and Schmidt 2006b], F&S use a more careful phrasing, saying that the model offers a *'unified interpretation',* and that the model's *major predictions* are consistent with the observed *qualitative pattern* of contract choice, but they hasten to add (on p.4) that the model also makes accurate quantitative predictions, (I describe these accurate predictions on p.14).

B.3 Some Properties of the $40-60$ Distribution

The new calibration is clearly *a set* of distributions (since the α, β are not specified), but F&S do not explicitly refer to it as a set, and they treat it as a single distribution. In all of the papers it seems that F&S' aim is to show that

all the distributions of the $40-60$ type are compatible with the experimental data.

The $40-60$ distributions preserve two properties of the QJE calibration. Like the QJE calibration it has 40% individuals with high $\beta's$, and like its predecessor it maintains a correlation between the two variables, thus the individuals with high $\beta's$ also have high $\alpha's$. The first property was presumably extracted from the data on ultimatum games (although the data was not provided), and the second property was introduced in an appendix of [Fehr and Schmidt 1999] as a tribute to reciprocity and an attempt to explain the Fehr and Gächter public good experiment.

In fact, although F&S claim in their papers to use the $40-60$ calibrations without any restrictions on the parameters, their proofs hold for a narrower set of calibrations. In [Fehr, Kremhelmer, and Schmidt 2005] F&S fix all the higher $\alpha's$ to be $\alpha = 2$, this, they claim, is *consistent* with their QJE paper. The value $\alpha = 2$ is *not* one of the values in the QJE calibration. F&S do not explain what the meaning of *'consistent'* is in this context.

In the other two papers, F&S change the calibration in the unpublished appendix, without informing the reader that the proofs do not apply to the general $40-60$ calibration. In the appendix to [Fehr, Klein, and Schmidt 2006b], F&S change all the higher values of the parameters to be $\alpha = 2, \beta = 0.6$. They do not even claim that this can be done without loss of generality.

The full extent of F&S' somewhat unorthodox methods is fully revealed in [Fehr and Schmidt 2004a] and in its unpublished appendix [Fehr and Schmidt 2004c]. In the text (p.470) F&S define individuals with $\alpha, \beta \geq 0.5$ as inequity averse individuals, and those with $\alpha, \beta < 0.5$ as selfish. They then declare that 'following' the QJE calibration there are 40% inequity averse individuals and 60% selfish ones. Whatever the meaning of 'following' in this context may be, this statement is false since the QJE calibration has 30% individuals who do not belong to either of the two categories (individuals with $\alpha = 0.5$, $\beta = 0.25$). The reader may not be aware that by 'following' the QJE calibration some 30% of its individuals were eliminated.

In the first page of the *unpublished* appendix [Fehr and Schmidt 2004c], where all proofs are to be found, F&S redefine selfish individuals to have $\alpha = \beta = 0$, (i.e. no longer $\alpha, \beta < 0.5$). This change is not mentioned in the main text of the paper, and all proofs in the appendix apply to this new version.

B.4 Incompatibility of the $40-60$ Distribution with the Experiments in the QJE paper.

The $40-60$ distribution is not compatible with the experiments in the QJE article (with the exception of the market experiment with proposers' competition, where the competitive outcome is independent of the population).

Compatibility with the market experiment with responders' competition requires that there be enough individuals with $\beta < 0.8\bar{3}$, this is not guaranteed for *all* $40-60$ distributions.

For the public good games without punishment, the theory with the $40-60$ population predicts that about 100% make no contributions, this is incompatible with the data, in which only 73% made no contributions.

The theory is incompatible with the Fehr and Gächter experiment. Proposition 5 requires that all individuals with $\beta > 0$ have $\beta \geq 0.6$ (and correspondingly high values of $\alpha's$). This is not guaranteed for *all* of the $40-60$ distributions.

In addition, the $40-60$ distribution is incompatible with the data on Ultimatum Games, which F&S used to calibrate their model. In the experimental Ultimatum Game data, as presented in QJE, 30% of the individuals have intermediate inequity aversion parameters.

B.5 Playing Down the Differences Between the $40-60$ Distribution and the QJE Calibration.

The two calibrations, that of QJE and the new $40-60$ one, are incompatible, as neither explains the experiments that the other was designed to explain. Yet, F&S do not mention this incompatibility in their papers. They do not tell their readers why they have switched to the new distribution, nor do they explain the significance of this change.

In fact, they do not even present the new distribution as a new one. They downplay the move to the $40-60$ distribution and present the new distribution as *following* the QJE calibration, as being *in accordance with*, or as *an aggregated and simplified version* of the QJE calibration ([Fehr and Schmidt 2004a], p.470, [Fehr, Klein, and Schmidt 2006b],p.22, and footnote 15, [Fehr, Kremhelmer, and Schmidt 2005], p.22,).

F&S do not explain what it means for two completely different distributions to be *'in accordance'* with each other, nor why it is permitted in this case to *'aggregate and simplify'* or to *'follow'* a distribution

B.6 The Relevance of the $40-60$ Distribution for Testing the Theory.

Although F&S disregard the quantitative predictions of the $40-60$ calibration they make the reader believe that this distribution is relevant for testing the the contract experiments.

Recall that F&S have publicly declared that in the contract papers, the inequity aversion theory produces *quantitatively accurate predictions*. These quantitative predictions must be those of the theory calibrated with the $40-60$ distribution which is used in these papers.

In the contract papers F&S repeat their claim that the QJE calibration is successful in explaining experiments accurately. They also present the $40-60$ distribution as a minor and innocuous variation of the QJE calibration which is suitable for the contract papers (see appendix B.5). The reader may be justified in believing that the $40-60$ distribution is relevant for explaining the contract experiments.

For the statements about the success of the QJE calibration see [Fehr and Schmidt 2004a], p.470, [Fehr, Klein, and Schmidt 2006b], p.21 and footnote 16, and [Fehr, Kremhelmer, and Schmidt 2005], p.20, 23 and footnote 18.

Any remaining doubt about the relevance of the $40-60$ distribution is removed when reading F&S' repeated proclamation in [Fehr, Klein, and Schmidt 2006b] that the model 'offers remarkably precise *quantitative* predictions of the data'. The quantitative predictions they refer to in these statements are clearly based on the $40-60$ distribution, (I describe these predictions on p.14).

B.7 A Larger Set of Distributions

In two of the contract papers F&S mention in passing that the qualitative properties hold for a wider set of distributions. Like the $40-60$ distributions, these are distributions with two types of individuals: the selfish and the fair ones but with different proportions. Thus, these distributions preserve the perfect correlation between the α, β parameters but allow deviations from the assumption of 40% fair individuals.

In [Fehr and Schmidt 2004a], p.470, F&S state that "The qualitative results are robust to changes in this distribution as long as there is a significant fraction of both inequity-averse and selfish players." No detailed range for these fractions can be found in the paper nor in its unpublished appendix.

In [Fehr, Klein, and Schmidt 2006b] p.25, F&S state that "... the qualitative results that follow are robust to changes in this distribution, as long as the share of fair types is at least 33 percent but not larger than 60 percent". There is no direct proof in the paper (or its appendix) that this is the relevant range, the figure 33 does not appear in the proofs.

In [Fehr, Kremhelmer, and Schmidt 2005] F&S do not claim that the propositions hold for a wider range of distributions. Indeed, a small change in the distribution renders their equilibrium invalid. A small increase, of 1.18%, in the group of fair individuals (i.e. the distribution becomes $41.18-58.82$ instead of $40-60$) may cause some fair agents (with $\alpha's$ close to 0.5) to behave like selfish agents.

B.8 The Robustness of the Equilibrium

F&S consider equilibria in which all the individuals of a certain type behave identically: Selfish agents expend high effort, fair ones shirk, fair principals pay selfish agents a bonus and selfish principals pay no bonus. To support these equilibria, the frequencies of the fair and selfish players in the population should be such that the selfish players would wish to cooperate and that all the fair agents would wish to 'shirk'. The equilibrium breaks down if the percentage of fair principals in the population is too low to support the cooperation of the selfish agents, it may also break down if there are too many fair principals, this may induce some of the fair agents to cooperate rather than shirk.

The parameters of the contract games (the production and cost functions) were carefully selected to ensure that in equilibrium all the fair agents play the same strategy, irrespective of their unspecified inequity aversion parameters. The production and cost functions guarantee that an agent expending high effort receives in the expected allocation (after the bonus), less than a quarter of the whole cake. This, in turn, guarantees that all the fair agents (with $\alpha > 0.5$) reject this allocation and choose a less efficient but more egalitarian allocation. However, the parameters of the game can be easily altered so that in the expected allocation the agent's share is more than a $\frac{1}{4}$ of the whole cake. In that case, some of the fair agents with $\alpha \sim 0.5$ may be induced to cooperate. The principals are, of course, interested to get the cooperation of more agents and will choose the wage appropriately. F&S do not tell their readers that the equilibrium is not robust to changes in the game parameters.

If equilibrium fails because of partial cooperation of the fair agents then in order to obtain a new equilibrium it will be necessary to assume a calibration

which has more than two types. The new calibration would either violate the perfect correlation between α, β by including individuals with high $\beta's$ but a low α - as principals they would be fair and as agents selfish, or else the calibration will have to detail the distribution of $\alpha's$ among the fair agents, so as to enable the principals to induce some fair agents to cooperate. In any case, a distribution with two types of individuals (like the $40-60$ distribution) will no longer suffice to describe the equilibrium behavior. F&S will have to change their calibration yet again. It also means that F&S will need to compute equilibria for populations with more than two types, which they have not done so far.

In a seminar Klaus Schmidt gave in Bonn in October 2005 (in the presence of Professors C. Engel, M. Hellwig and U. Schweizer), he admitted that he tried to apply the QJE calibration to the Fehr-Klein-Schmidt model but failed to find an equilibrium because of the multiple *(four)* types in the calibration.

B.9 The Contract Experiments: Testing the Theory

We use the following method for the tests: we assume that the empirical behavior follows the suggested equilibrium and estimate the percentages of fair principals and fair agents in the population. The fraction of agents who expend low effort is an estimate for the fraction of fair agents, similarly, the fraction of fair principals can be estimated by computing the fraction of principals who rewarded agents who exerted high effort. We then check whether the estimated population agrees with the calibrations.

When we compare the theory's predictions with the data we find large disagreements between the two. The fractions of fair agents (those who make a low effort) and of fair principals (those who reward high effort) that we find in the data are inconsistent with the calibrations considered by F&S.

If there is anything to be learnt from the data it is that there is no perfect correlation between the parameters α, β and that no single calibration can explain the different experiments. The attempt to explain all three experiments with a distribution similar to the $40-60$ distribution has failed.

In two of the experiments ([Fehr and Schmidt 2004a], [Fehr, Kremhelmer, and Schmidt 2005]), the data suggests that the number of fair principals (who pay the bonus) is much larger than the number of fair agents (who shirk). In the first paper the number of fair principals is 3 times that of the fair agents, and in the second paper it is 13 times. This is incompatible with the distribution which F&S assume to prove their propositions. The different proportions of fair agents to fair principals in the two experiments suggests that the calibrations required to explain the data will differ drastically between the experiments.

Those few individuals who shirked despite the almost certain bonus, must have very large $\alpha's$. If F&S were to look for a calibration that describes the data they will have to assume a particular distribution of $\alpha's$ among the fair individuals. It seems, that in order to explain the data it is required either to relax the perfect correlation between α, β (by introducing individuals with high $\beta's$ but low $\alpha's$), or else have more information about the distribution of $\alpha's$ among the fair individuals. In any case the calibration will no longer be the $40-60$ distribution, and it will have to have more than the two types assumed by F&S..

Moreover, in [Fehr and Schmidt 2004a], the empirical average payoff of an agent is double the predicted payoff, bonuses are not paid to equalize the payoffs of the agent and the principal, and at least 25% of the agents do not act according to the calibrated theory.

In [Fehr, Klein, and Schmidt 2006b], the proportions are reversed, the data suggests that there are more fair agents than fair principals. The low fraction of fair principals suggested by the data does not support any cooperation in the theoretical model. There seems to be more cooperation in the experiment than the theory can explain. In addition, the bonuses paid in the experiment do not equalize the payoffs of the principal and the agent, whereas the theory assumes that bonuses are paid to achieve equality.

We now describe the tests for each of the papers:

Fehr and Schmidt 2004a

In [Fehr and Schmidt 2004a], according to the equilibrium of the bonus contract, all principals pay a wage of $w = 225$, selfish agents expend (total) high effort $e = 20$, and are paid by the fair principals a bonus of 350. Fair agents expend a (total) effort of $e = 12$, and are paid no bonus. Proposition 2 ([Fehr and Schmidt 2004c], p.$A-10$) lists all pooling equilibria for any permissible fractions q of fair players in the population, Proposition 3, selects one of these with the help of a refinement-like argument, (Condition 1 in p.$A-11$).

We find that the data is incompatible with the theory in many ways. The theory predicts that the fair agents make a low total effort of $e = 12$, taking the range of (total) effort levels $10-14$, we find that the fraction of those who made efforts in this range is: $55/261 = 21\%$, (Table 2, p.463). This suggests that the percentage of fair agents is 21%. Taking the interval $[18, 20]$ for the high effort level $e = 20$, the percentage of those who made an effort ≥ 18 is $93/261 = 35.6\%$, suggesting that the percentage of selfish agents is 35.6%. This leaves 43.4% of agents who do not behave according to the theory. Indeed, about 25.6% of the agents made total efforts $e = 2, 3, 4$, these effort levels are incompatible with the theory.

There is no detailed data on the paid bonuses in the paper, we therefore use the following method to estimate the percentage of fair principals in the population. The average bonus paid for the high effort levels $18, 19, 20$ is about 211 (Table 2 on p.463 and figure 3 on p.464): $\frac{24}{24+6+63} * 120 + \frac{6}{24+6+63} * 170 + \frac{63}{24+6+63} * 250 = 211.29$. A fair principal pays a bonus of 350 in the equilibrium, the estimated fraction of fair principals in the data is therefore: $\frac{211}{350} = 60.3\%$. This is about 3 times the estimated percentage of fair agents (21%) . This is incompatible with the calibrations assumed by F&S.

We also compare the theoretical and empirical average payoff of an agent. The empirical average payoff of an agent is about 400, (according to figure 4 on p. 467). The theoretical average payoff of an agent is:

$$[75 + 350q](1 - q) + 155q,$$

where q is the fraction of fair players, 75 is the equilibrium pre-bonus payoff of a selfish agent, and 155 is the payoff of a fair agent. The maximal value of this function is: 207. The maximum possible theoretical average payoff of an agent is about half the empirical one.

According to the theory bonuses are paid to equalize the payoffs of the principal and the agent, we test whether the data confirms it. Let v, c be the production and cost functions (of effort), and w, b be the wage and bonus paid. Equating the payoffs of the principal and agent implies: $b = \frac{v+c}{2} - w$. Given the frequencies of pairs of efforts in the experiment (Table 2, p.463), we can calculate the average $\frac{v+c}{2}$ for any range of efforts. There is no information in the paper about the wage paid, I will, therefore, assume it to be the equilibrium wage $w = 225$. Thus we can calculate the average bonus which *would have* equated payoffs, for a certain range of efforts. Using Table 2 (frequencies of efforts, p.463) and Figure 3 (average bonus paid per effort level, p. 464) we can compute the average bonus that was actually paid for this range of efforts. We can then compare the two.

The effort range that I consider is $e_1 + e_2 \geq 13$, since for total effort ≤ 12, and for $w = 225$, the principal's payoff is lower than the agent's and he cannot pay a bonus. For all the cases of total effort $e_1 + e_2 \geq 13$ the principal can pay a bonus.

For the range of total efforts $e_1 + e_2 \geq 13$ the average bonus (actually) paid is about 180.73, while the average bonus required to equalize the payoffs is 245.77, the discrepancy between the two is higher than 35% $\left(\frac{245.77-180.73}{180.73}\right). = 35.98\%$ If we allow the bonuses not to equate the payoffs but to set the agent's payoff at 80% of the principal's, the gap reduces to about 15%. Clearly the bonuses were *not* paid in order to equalize payoffs.

There were 60 subjects in the relevant part of this experiment and 257 observations.

Fehr, Klein and Schmidt 2006 (Forthcoming in Econometrica)

In the equilibria of [Fehr, Klein, and Schmidt 2006b], with the calibration considered by F&S, all principals pay a wage $w = 15$, selfish agents make an effort of 7 and the fair principals pay them a bonus of 25, fair agents make a low effort $e = 2$ and no principal pays them a bonus.

Table V in the paper, describes the bonus to effort relation under bonus contract in two sessions of the experiment $(S3 - S4)$. According to this table, low efforts ≤ 3 were made in $\frac{55}{198} = 27.7\%$ of the cases, suggesting that the percentage of fair agents is 27.7%. High efforts ≥ 5 were expended in 127 of the cases. Of those who made high efforts, the fraction of those who received a bonus ≥ 21 is $\frac{36}{127} = 28.3\%$. This suggests that the fraction of fair principals (who pay the bonus) is 28.3%. The numbers of fair agents and principals are about the same.

A word on the intervals I chose to represent the equilibrium values: the effort scale is rather coarse: the integers $1 - 10$. I therefore allowed an individual who chooses an effort level $e = 2$, to err 50% and choose an effort in the interval $[1, 3]$. For the equilibrium wage $w = 15$, the effort level $e = 1$ results in a very unequal allocation:$(15, -5)$, yet, I take it to be a proxy for $e = 2$. If it were not the fair agents who chose $e = 1$, then we are left with 24% of the agents $\left(\frac{91}{376} = 24.2\%\right)$ who choose an effort level which is incompatible with the theory.

I apply a similar argument for taking the efforts interval $[5, 10]$ to represent high efforts $(e = 7)$, due to the coarsness of the effort scale, I allow the agent to err 28% downwards and 43% upwards.

Bonuses were selected from a finer scale, the integers $1-40$. The theory predicts that when the agent made the effort $e=7$ and was paid a wage $w=15$, the bonus paid should equate the payoffs of an agent and the principal. A bonus of 18 (the average of the range $16-20$) results in an unequal allocation, it leads to an allocation $(15-10+18, 70-15-18)=(23,37)$ in which the agent receives only 62% of the principals's share. I therefore consider the bonuses in the interval $[21,40]$.

F&S have pooled the data of all the bonus contract games of the various treatments (sessions $S3-S6$) in Table 5 of the unpublished appendix [Fehr, Klein, and Schmidt 2006a]. They inform us that there is no statistically significant difference in the bonus-effort relation between the two tables (Result $6\,(b)$, p.19). Indeed, we run the same test for the fractions of fair principals and fair agents on the pooled data, and obtain similar results with one major difference. The fraction of fair principals is 23.9% $\left(=\frac{51}{213}\right)$, but the fraction of fair agents is much higher: 34.5% $\left(=\frac{130}{376}\right)$, i.e. there seem to be significantly more fair agents than fair principals (about 44% more). This is incompatible with F&S's assumption of perfect correlation between α, β.

The fact that the data of sessions $S3-S4$ suggests a different composition of population than the data for sessions $S3-S6$ shows that there is little hope to find a *single* calibration that will explain all experiments. In all the sessions the bonus contract game was played, the sessions differ in their framing and in the number of contract types that the principals may select from.

Both tables agree that the percentage of fair principals is low, between 23.9 and 28.3%. According to F&S the equilibrium they consider (or rather its *'qualitative results'* in their formulation) holds when the percentage of fair principals is not lower than 33%, but they make no claims about the existence of equilibrium for lower percentages. We show, under very weak assumptions, that for any q, and any calibration with a fraction q of fair principals (high $\beta's$), the data is incompatible with this calibration. This means that there is more agents' cooperation in the experiment than the theory (as presented by F&S) can possibly support. (Note that we make no further assumptions about the population, except assuming that there is a fraction q of players with $\beta > 0.5$).

For a given q, we find the effort level e^S that a selfish agent will expend when he believes that a fraction q of the principals will reward him. This level is a function of q only and not of the wage paid by the principal (see lemma 2, [Fehr, Klein, and Schmidt 2006a]). The total compensation paid by a fair principal to an agent who made the effort e^S is $b+w=\frac{v\left(e^S\right)+c\left(e^S\right)}{2}$, where v, c are the production and cost functions, respectively The theory predicts that $b+w$ should be paid by the q fair principals to those agents who made the effort e^S. F&S inform us that the wage w was close to 15 throughout the experiment. Using Table V, we can, therefore, find out what percentage of the principals paid a bonus of $\left[\frac{v\left(e^S\right)+c\left(e^S\right)}{2}-15\right]$ to those agent who made the effort e^S. Then, we can compare it with the assumed fraction q. We find that the two values never agree.

The calculation is rather tedious and is presented below.

The function $e^S(q)$ is a step function, because the cost function is a piecewise linear function.

- For $q < 0.1818$, the selfish agent will choose $e = 1$, there is no equilibrium with cooperation.

- For all $q's$ in the interval $[0.1818, 0.33)$ the selfish agent will choose $e = 3$. The bonus is: $b = \frac{10*3+2}{2} - 15 = 1$. Out of 156 who chose $e \geq 3$, about 127 were paid bonuses $b \geq 1$, suggesting that the fraction of fair principals is $\frac{127}{156} = 0.814$. This value is outside the relevant interval of $q's$. The data is incompatible with calibrations chosen in this range of $q's$. But there is more incompatibility: for $w = 15$ (the actual wage paid) there cannot be any cooperation in this range, the selfish agents will choose not to cooperate (in the equilibrium the wage should be ≤ 9.9), however, in the experiment about 50% of the agents exerted high efforts ≥ 6 .

  This range of $q's$ is an example of how sensitive the analysis is to the parameters. If we follow F&S and the data, and take a two type calibration with about 28% of fair and 72% of selfish players, we will need to know more about the distribution of α, β among the fair players in order to compute the actions of the fair agents. The effort exerted by a fair agent (with both $\alpha, \beta > 0.5$) depends on the exact values of α, β. He will choose $e = 1$, or $e = 2$ depending on whether α is close to β or much larger. This, in turn, affects the wage chosen by the principal.

  With further assumptions on the populations we may succeed in finding an equilibrium which will have a low degree of cooperation, the agents' efforts will all be ≤ 3. This makes the bonus contract rather unattractive, and it now becomes questionable whether the principals will choose the bonus contract in their first move. Depending on the cost of practicing punishments (verification costs) in the incentive contracts, the principals may find the incentive contract more attractive, contrary to the behavior in the experiment.

- For all $q's$ in $[0.33, 0.46)$ the selfish agent chooses $e = 7$. The bonus is: $b = \frac{10*7+10}{2} - 15 = 25$. We test for $e \geq 5, b \geq 21$, and find that the estimated fraction of fair principals is $\frac{36}{127} = 0.283$, outside this range.

- For all $q's$ in $[0.46, 0.571)$ the selfish agent chooses $e = 9$. The bonus is: $b = \frac{10*9+16}{2} - 15 = 38$. We take the range $e \geq 7, b \geq 31$. Out of 82 who exert high efforts 9 were given bonuses, the estimated fraction of fair principals $\frac{9}{82} = 0.109$, is well below the relevant range of $q's$.

- For all $q \geq 0.571$ the selfish agent chooses $e = 10$. The bonus is: $b = \frac{10*10+20}{2} - 15 = 45$. We test for $e \geq 8, b \geq 36$ and find that out of 46, only 6 were paid bonuses, $\frac{6}{46} = 0.13$, which is not in the relevant range.

It follows that there is no q and no calibration with q fair agents that is compatible with the data of this experiment. The only equilibrium for a population (with two types, as considered by F&S) and with a low percentage of fair principals (as suggested by the data) will have very little cooperation, all agents will expend low efforts.

The theory predicts (or rather assumes) that positive bonuses are paid in order to equate the payoffs of the agent and the principal. We can test whether the data confirms it. To perform this test we need to know the empirical wage

paid to the agent, F&S do not provide detailed information about the wage paid, they provide only the average wages as a function of time (figure 5). To test whether the principals attempted to equate payoffs via the bonuses, I assume that the wage paid was $w = 15$, this is the theoretical wage and it also happens to be the average empirical wage. I compute the fraction of bonuses which give the agent at least 80% of the principal's share, these are the bonuses for which $b + w \geq \frac{0.8v+c}{1.8}$.

We consider only bonuses ≥ 6, (assuming that those principals who paid a bonus below 6, are selfish and meant to pay 0). We find that of the bonuses ≥ 6, only $\frac{44}{106} = 41.5\%$ were close to equating the payoffs of the agent and the principal, whereas the theory predicts that *all* positive bonuses equate the payoffs. For the pooled data of Table 5 of the unpublished appendix [Fehr, Klein, and Schmidt 2006a], the fraction of the bonuses which roughly equate the payoffs is 43.7%. Clearly, the bonuses paid in the experiment are not meant to equate the payoff of the principal and the agent.

F&S claim that the bonuses paid in the experiment form a substantial part of the agent's compensation. They calculate the average bonus paid in the experiment (10.4) and the average wage (15) . The bonus part of the total compensation to the agent is: $\frac{10.4}{25.4} = 40.9\%$ We now compute this ratio for the equilibrium (of the $40-60$ calibration). The average bonus is $25 * 0.4 * 0.6 = 6$, the wage is 15, hence the theoretical bonus part of the agent's total compensation is: $\frac{6}{15+6} = 28.5\%$. This is much lower than the experimental value, 40.9%. Again, there seems to be more cooperation in the experiment than the theory permits.

We briefly compare the equilibrium with the experimental behavior under an incentive contract. According to the theory, (Proposition 2, p. $A-3$, [Fehr, Klein, and Schmidt 2006a]), all principals demand an effort level of $e^* = 4$, the selfish principals offer a wage of $w = 4$ and the fair ones $w = 17$. All agents accept the fair offer and fair agents reject the selfish offer. Comparing this with the data (the first part of Table III) the percentage of fair principals, who made a high wage offer is $26/56 = 46.4\%$. The percentage of fair agents, those who reject a low wage offer is: $8/26 = 30.7\%$, substantially lower than the fraction of fair principals. In addition, the average payoff of a principal does not match. In the experiment it is 8.6. According to the theory, the average payoff is a weighted average of 15.6 and 13 : $15.6(1-q) + 13q$ where q is the fraction of fair principals, this is well above 8.6 for all $q's$.

There were 88 subjects in this experiment and 376 observations.

Fehr, Kremhelmer and Schmidt 2005

The case of [Fehr, Kremhelmer, and Schmidt 2005] is straightforward. Here the propositions were proved for the $40-60$ distribution, F&S do not claim that they hold for a wider set of distributions. Indeed, a small increase in the percentage of fair principals can destroy the equilibrium, (increasing the percentage of fair principals to 41.2% will make a fair agent with α close to 0.5 cooperate like a selfish agent).

For this game the theory predicts (propositions $3, 4$ pp. $25, 26$) that a selfish B player makes the effort $b = 10$, and a fair one sets $b = 1$. A selfish A player chooses the effort $a = 1$ and a fair A sets his a to equal b.

According to Table 3 on p.17, the fraction of B agents who chose low effort levels $1, 2, 3$ is $\frac{10}{187} = 5.34\%$, suggesting that the percentage of fair agents is 5.34%.

Of those B players who expended high efforts ($b = 8, 9, 10$), a fraction $108/155 = 69.67\%$ were rewarded by a high effort of the A player, this suggests that the percentage of fair principals is 69.67%.

There is a huge gap between the percentages of fair agents and fair principals. This is incompatible with F&S' calibration. Moreover, this high percentage of fair principals (69.67%) cannot support the equilibrium considered by F&S, if the percentage of fair principals is higher than 41.2 then some fair agents (depending on their α values) may cooperate.

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

There were 44 subjects in this treatment and there were 187 observations. The above computations do not change significantly when we combine this test with the JOD treatment (Table 2 p.13), the combined sample size is 487 with 132 subjects.

B.10 Rhetorical Devices

The danger of embarking on a project without a clear methodological plan becomes apparent when we consider the entirety of the papers. The whole project is shrouded in mist. F&S choose their calibration by using data they want to explain, but hide this fact from the reader. They manipulate and modify the calibration in an appendix but do not mention it in the main text of the paper. They prove a theorem that does not apply to their calibration but claim that it explains the data.

In the contract papers, F&S change the calibration, but do not discuss the problematic nature of this change. Instead, they use vague, ambiguous wording to hide the change. In their unpublished appendices they manipulate the calibration yet again, without informing the readers that the proofs apply only to the manipulated version. The new calibration provides detailed predictions which do not agree with the experimental data. F&S do not even attempt to test these predictions, they simply ignore them without giving any reason for it. The predictions strongly disagree with the experimental data, yet on the basis of some rudimentary features F&S claim that the theory largely fits the data. When the data accidentally agrees with the theory on some averages, although the underlying distributions are wide apart, F&S repeatedly hail this find as an 'accurate quantitative prediction'.

Judging from the methods they use in the contract papers, F&S no longer believe in the existence of a single calibration that can explain all their experiments. Yet, they use vague phrases so as to make the reader believe that the project still runs in its announced course.

All this does not prevent F&S from claiming, in the public discussion of their theory on the internet, in survey papers written for distinguished societies and in learned handbooks that they have kept their calibration constant and that their calibrated theory yields accurate predictions of the data.

References

Fehr, E., and S. Gächter (2000): "Cooperation and Punishment in Public Goods Experiments," *American Economic Review*, pp. 980 – 994.

Fehr, E., A. Klein, and K. M. Schmidt (2006a): "Appendix to 'Fairness and Contract Design'," http://www.vwl.uni-muenchen.de/ls_schmidt/who/personen/schmidt-papers/Fehr-Klein-Schmidt-Appendix(2006).pdf.

——— (2006b): "Fairness and Contract Design," mimeo, http://www.iew.unizh.ch/home/fehr/downloads/MS5182-Fairness_and_Contract_Design_ECMT_Final.pdf, forthcoming in Econometrica.

Fehr, E., S. Kremhelmer, and K. M. Schmidt (2005): "Fairness and Optimal Allocation of Property Rights," Discussion Paper no. 5369, CEPR, London.

Fehr, E., and K. M. Schmidt (1999): "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114, 817 – 868.

——— (2003): "Theories of Fairness and Reciprocity: Evidence and Economic Applications," in *Advances in Economic Theory, Eigth World Congress of the Econometric Society*, ed. by S. T. M. Dewatripont, L.P. Hansen, vol. Vol. 1, pp. 208–257. Cambridge University Press, Cambridge, Vol. 1.

——— (2004a): "Fairness and Incentives in a Multi-Task Principal-Agent Model," *Scand. J. of Economics*, 106(3), 453 – 474.

——— (2004b): "The Role of Equality, Efficiency, and Rawlsian Motives in Social Preferences: A Reply to Engelmann and Strobel," Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 179.

——— (2004c): "Theoretical Appendix to 'Fairness and Incentives in a Multi-Task Principal-Agent Model'," http://www.vwl.uni-muenchen.de/ls_schmidt/experiments/multi_task/index.htm.

——— (2005): "The Rhetoric of Inequity Aversion - Reply," http://www.vwl.uni-muenchen.de/ls_schmidt/pamphlet/Shaked-Reply.pdf.

——— (2006): "The Economics of Fairness, Reciprocity and Altruism Experimental Evidence and New Theories," in *Handbook of the Economics of Giving, Altruism and Reciprocity, Volume*, ed. by S.-C. Kolm, and J. M. Ythier, chap. 8. Elsevier B.V., DOI: 10.1016/S1574-0714(06)01008-6.

Fischbacher, U., C. Fong, and E. Fehr (2003): "Fairness, Errors and the Power of Competition," Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 133.

Gintis, H., S. Bowles, R. Boyd, and E. Fehr (2003): "Explaining Altruistic Behavior in Humans," *Evolution and Human Behavior*, 24, 153 – 172.

Güth, W., N. Marchand, and J.-L. Rulliere (1997): "Ultimatum Bargaining Behavior - A Survey and Comparison of Experimental Results," Discussion Paper, Humboldt University, Berlin.

Roth, A. E. (1995): "Bargaining Experiments," in *Handbook of Experimental Economics*, ed. by J. Kagel, and A. Roth. Princeton University Press.

Roth, A. E., V. Prasnikar, M. Okuno-Fujiwara, and S. Zamir (1991): "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review*, LXXXI, 1068 – 1095.

Shaked, A. (2005): "The Rhetoric of Inequity Aversion," http://www.wiwi.uni-bonn.de/shaked/rhetoric/Rhetoric-8-3-05.pdf.