ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

IS TRUST SELF-FULFILLING? AN EXPERIMENTAL STUDY

Michael Bacharach, Gerardo Guerra and Daniel J. Zizzo

Number 76

October 2001

Manor Road Building, Oxford OX1 3UQ

tr\dispap76.tex

IS TRUST SELF-FULFILLING? AN EXPERIMENTAL STUDY

Michael Bacharach, Gerardo Guerra and Daniel J. Zizzo [1]
Department of Economics, University of Oxford

October 2001

ABSTRACT

A person is said to be 'trust responsive' if she fulfils trust because she believes the truster trusts her. The experiment we report was designed to test for trust responsiveness and its robustness across payoff structures, and to disentangle it from other possible factors making for trustworthiness, including perceived kindness, perceived need, and inequality aversion. We elicit the truster's confidence that the trustee will fulfil, and the trustee's belief about the truster's confidence after the trustee receives evidence relevant to this. We find evidence of strong trust responsiveness. We also find that perceptions of kindness and of need increase trust responsiveness, and that perceptions of kindness and need raise fulfilling rates only in conjunction with trust responsiveness.

JEL Classification Numbers: C79, C92, D84. KEYWORDS: trust game, experiment, trust responsiveness, kindness, need to trust, belief elicitation.

[1] We would like to thank Abigail Barr, Jordi Brandts, Andrew Colman, Miguel Costa-Gomes, Jim Engle-Warnick, Diego Gambetta, Sean Hargreaves-Heap, David Hendry, Graham Loomes, Chris Starmer and Bob Sugden for criticisms and comments. We owe particular debts to Justin Smith and Bronwyn Hall, who made major contributions to the development of the experimental design and to the econometric analysis of the data, and to British Telecom Research Laboratories for financial support under agreement ML826341.

INTRODUCTION

Today few doubt the importance of trust and trustworthiness as explanatory factors in economic behavior. It is also held that they are fundamental to economic welfare: they allow saving on the costs of writing, policing and enforcing contracts, and are even preconditions for the existence of markets. They explain the prevalence of honesty in making social security claims, the custom in restaurants of serving first and charging afterwards, unmonitored time-based payment schemes, and the general acceptance of informal promises in trade. They constitute a good proportion of the 'social capital'. But despite the centrality of trust and trustworthiness in economic activity, and despite the widespread recognition today of their centrality, there remains much mystification about what produces them, and even about what trust is.

Like their cousin, cooperativeness, the T-pair – trust and trustworthiness – have proved hard to accommodate in the framework of rational decision theory (Hollis 1998). This has led some to denigrate them as irrational, however socially beneficial they might be. Others (Hardin 1991) have sought to rationalize the T-pair, typically as strategies in repeated interactions. Yet others have explained them as the product of motivational traits that are neither rational nor irrational (Bacharach and Gambetta 2001a). One example of this last view is the suggestion that trustworthiness can be produced by a motivational trait known as *trust responsiveness* or *the self-fulfilling property of trust*,[1] the tendency to fulfil trust because you believe it has been placed in you. The present paper is an investigation of this suggestion.

The question of the existence of trust responsiveness is of considerable practical importance. This paper grew out of research into the role of trust in e-commerce, where the lack of trust and trustworthiness is often seen as depriving society of large potential welfare gains. If trustworthiness is indeed produced by trust responsiveness, then trustworthiness in the world can be enhanced, and these welfare gains realized, by any signal that credibly informs the trustee of the truster's trust when this is present. This route to harvesting the gains to society from trust may be far easier, quicker and cheaper than the reform of deep-grained cultural attitudes by a long and radical process of re-education that is sometimes held to be needed (Putnam *et al.* 1993).

Before going any further we note, as Hausman (1998) does, that the key to the 'puzzle of trust' is likely to lie on the side of trustworthiness. Once it can be shown

that it is reasonable to expect trustworthiness there is no longer any mystery about trust, since trust is typically a best reply to this expectation. The puzzle is to explain why reasonable people should have the expectation, since in typical trust situations trustworthy behavior goes directly against material incentives.

In Section 1 we define a paradigm class of trust games, and within it the notions of trust and trustworthiness. Next we review the recent experimental literature, the hypothesis of trust responsiveness and some leading alternative theories of trustworthiness. In Section 2 we describe the design of an experiment. It aims to answer the question 'Does trust responsiveness exist?', and to determine how, if so, trust responsiveness relates to alternative explanations. The experimental methods are relatively new: instead of collecting purely 'behavioral' data and trying to infer underlying strategies and beliefs from this information alone, we elicit strategies and beliefs directly, with due attention to incentives, so obtaining a much richer body of data with which to test competing theories. In the present case they yield clear answers to key questions. Section 3 reports results, Section 4 discusses the formation of subjects' beliefs and the direction of causality between rates of fulfilment and other variables, and Section 5 summarizes and draws some implications for policy and theory.

1. TRUST AND TRUST RESPONSIVENESS

1.1 Trusting and Fulfilling

The most elementary kind of situation in which it is correct to speak of 'trust' is a two-person game, the Basic Trust Game, whose normal form is shown in Table 1, and whose coefficients satisfy the three inequalities:

(1) $y < a$ (Exposure),

(2) $a < w$ (Improvement),

(3) $x < z$ (Temptation).

The row player's two strategies are called Trust (T) and Withhold (W), and the column player's are called Fulfil (F) and Violate (V). The row player is called the *truster* (R)

Table 1: Basic Trust Game in Normal Form

	Trustee (E)	
Truster (R)	Fulfil (F)	Violate (V)
Trust (T)	w, x	y, z
Withhold (W)	a, b	a, b

and the column player the *trustee* (E). Inequality (1) implies that R is exposing herself to a risk by trusting, as she could be made worse off by it (she will be if E violates). Inequality (2) means that R can be made better off by trusting E (she will be if E fulfils). Finally, (3) means that E has an incentive to violate, his 'temptation'.

Normal form games with the payoff structure of the Basic Trust Game typically arise in the real world from extensive forms in which the truster R moves first and the trustee E observes her choice and responds. For example, it might be that R has to decide whether to inform E of a profitable opportunity from which R acting alone can make only 100, while E, who can make 200 from it, has the options of returning 150 and keeping 50, or returning nothing and keeping all 200: here $w = 150, x = 50, y = 0, z = 200, a = 100$, and b is unspecified.

Most writers on trust have agreed that the inequalities (1)-(3) are constitutive features of situations in which trust can occur. Some (Coleman 1990, Bacharach and Gambetta 2001a) go further, and use them to define such situations. Bacharach and Gambetta, for example, say that R trusts E to do X if she faces a BTG with $\mathsf{F} = X$ and chooses T because she expects E to choose F. It is common to regard the cooperative choice by the first mover in a 'staggered' Prisoner's Dilemma as a case of trusting (Wrightsman 1966, Hausman 1998), and so it is, on this analysis, if we take the second mover's strategy set to be 'C if and only if player 1 plays C' and 'D regardless'.[2]

According to some writers, however, a further inequality is also constitutive of trust situations, namely:

(4) $b < x$. (Mutual Gain)

If (4) holds then, in view of (2), the T-pair Pareto-improves on the *status quo*. There

are many important trust problems, including the staggered Dilemma, in which it does. But not all situations in which it is normal to speak of trust and trustworthiness satisfy (4): frequently, a person trusts another to behave in a way which makes him worse off than he was, as when an escaping prisoner of war trusts a cottager to shelter him at a risk to the cottager's own life.

1.2 Behavior in Trust Games

In any BTG, conventional game theory predicts (W, V): R is sure that E will play the weakly dominant strategy V, and there is no trusting. This is equally true in sequential versions of BTG and in 'fractional' versions in which fulfilment and trusting are matters of degree.[3] Yet in all forms of trust game which have been studied in the laboratory, the conventional game-theoretic prediction is massively violated.

The experimental literature on the BTG and its variants includes Berg *et al.* (1995), Bolle (1998), Dufwenberg and Gneezy (2000), Engle-Warnick and Slonim (2001), and Fehr and Gächter (1997). The central result is that half or more R subjects play T (or in fractional versions give half or more of their endowment), and many subjects fulfil to a substantial degree. For example, in Bolle's design if R transfers 80 DM to E, E receives double this amount, and E then plays a dictator subgame. Three quarters of trusters chose T, and the average sum returned was not significantly different from 80 DM. In Berg *et al.*'s design R could transfer any sum up to \$10, providing E with three times the transfer; R subjects transferred just over 50 percent on average, 90 percent of them a positive amount, and nearly 50 percent of E subjects returned more than the amount R transferred. Because these are not repeated games, trusting and fulfilling behavior cannot be explained as rational strategies for getting future rewards.

1.3 Theories of Fulfilling

Perhaps the commonest explanation in game theory of trusting and the fulfilling of trust is in terms of reputation and long-term reward or loss. But although such forces may well be at work when interactions are repeated many times, this approach can only explain the laboratory evidence about single or short interactions with strangers if it is combined with a strong form of 'assimilation' hypothesis.[4]

Another response has been 'transformed-payoff' theories of games. In this kind of theory each player i is ascribed a 'primary' payoff u_i, and an 'all-in' payoff U_i which is

a function of u_i and of further arguments. The primary payoff is typically the player's utility from her material reward. It is the all-in payoff U_i that determines i's choices. The transformed-payoff structure allows one to represent many possible psychological motivations (Zizzo 2000). For example, a utilitarian altruist i can be represented as having the payoff $U_i = \sum_j u_j$.

Formal theories of trust responsiveness represent the disposition by a special kind of payoff transformation. A 'psychological game' (Geanakoplos *et al* 1989) is a game with transformed payoffs in which a player's secondary utility is a function of her belief about players' beliefs about players' choices. The hypothesis that people are trust responsive may be represented in this way by using the link between trusting and expecting. Consider the psychological game in which primary payoffs are as in Table 1, R's all-in payoff is just her primary payoff, but E also has a utility from doing F if she believes that R believes she will do F. Then E is trust responsive, that is, has a preference for F if he believes that R *trusts* him, for if he believes this he also believes that R expects him to play F,[5] and then playing F gives him secondary utility. In this paper too we shall represent trust responsiveness in this style, in terms of second order belief.

Unlike many transformed-payoff theories, the trust-responsiveness theory and the alternatives to it we investigate here do not embody an equilibrium assumption. There are two reasons for this. First, in brief interactions betwen strangers the case for expecting equilibrium behavior is weak. Secondly, the particular transformed-payoff theories we are interested in postulate explicit relationships between preferences and beliefs. The equilibrium assumption is often needed to render such theories testable, which it does by eliminating beliefs. It is not needed when, as in our design, one gathers direct evidence on players' beliefs.

Allowing payoff-transformations into the theory of games creates an identification problem. Very often the same behavior is predicted by more than one perfectly plausible transformed-payoff theory. For example, cooperation in a sequential Prisoner's Dilemma could be the effect either of reciprocity or of altruism. This is a serious difficulty, but it is one that the laboratory can sometimes overcome. The experiment we report was expressly designed to test for the presence of trust responsiveness in a way which does not confound it with alternative sources of secondary payoff from fulfilment.

1.4 Trust Responsiveness

When someone lends someone money, or leaves the children in charge of the house, or holds an uninvigilated exam, she trusts others. And then it is quite common for the truster to say "I'm trusting you to ... " or "I know I can trust you to ... ". When she does, she feels that her message, if believed, will improve her chances that her trust will be fulfilled. If her trustee is trust responsive, she is right.

Henceforth we write t for the probability with which the truster R chooses T and f the probability with which the trustee E chooses F. We let t^* denote E's estimate of t, f^* R's estimate of f, and f^{**} E's estimate of f^*. We call f the trustee's *propensity to fulfil*, f^* the truster's *confidence*, and f^{**} the trustee's *confidence-perception*.

Trust responsiveness is the effect on the trustee's propensity to fulfil of her confidence-perception. Trust responsiveness implies that f increases with f^{**}. But this is not quite enough to characterize the intuitive notion: we must add the proviso that the function expresses a causal relation from f^{**} to f; E must be made more ready to play F *because* she believes that R expects her to. As we shall see, there are other possible patterns of causality which might surface in a positive association between f and f^{**}. In sum, a trustee is *trust responsive* if an increase in f^{**} tends to bring about an increase in f.

Numerous authors through the centuries have conjectured and discussed trust responsiveness (Bacharach and Gambetta 2000b, Gambetta 1988, Hausman 1998, Hirschman 1988, Hume 1740, Jussim 1986, Pettit 1995). But what lies behind it is anything but obvious. Two elements in the informal explanations of trust responsiveness in the literature are aversion by the trustee to 'letting down' the truster, and the idea that this aversion depends on the sympathy or respect the trustee feels for the truster – on how 'pro' his attitude towards her is. The aversion to letting down suggested by Dufwenberg and Gneezy (2000) could have two principal sources. First, 'outcome disappointment' on R's part. If R expects the good outcome (T,F), in which she gets w, she will be disappointed by the bad outcome (T,V), and this disappointment may be increasing in her *ex ante* confidence f^*. Suppose all this is in E's model of R. Then if E has sympathy for R he will have secondary utility from (T,F) which decreases in f^{**}. Second, 'person disappointment' on R's part. The trustee might be concerned about disappointing R's expectations not about her payoff but about him as a trustworthy person. He may value the good opinion of others (Hume 1740), especially those he respects (Hausman

1998).[6] Hausman adds that the more certain someone is of this good opinion, the more strongly he will wish to keep it. Then if E respects R, the surer he is that R thinks him trustworthy the more he will wish to fulfil trust – he will have secondary utility from (T, F) increasing in f^{**}.

If, as this analysis suggest, the motives underlying trust responsiveness depend on the sympathy or respect the trustee feels for the truster, then the strength of trust responsiveness may vary with aspects of the payoff structure which promote or discourage these attitudes in the trustee. Our experiment design affords a test of this postulated feature of trust responsiveness.

1.5 Other Transformed-Payoff Theories of Fulfilment

Trustworthiness can be explained by *inequality aversion* in the trustee. If E is inequality averse in the sense of Fehr and Schmidt (1999), his all-in payoff is $V = v + \psi$, where v is his primary payoff, his secondary payoff is $\psi = -\alpha(v-u)$ if $v \geq u$ and $= -\beta(u-v)$ if $v \geq u$, where u is R's primary payoff, and α, β are personal parameters with $0 < \alpha < \beta$. Thus E's all-in payoff gain from choosing F rather than V when R chooses T is

$$V(\mathsf{T},\mathsf{F}) - V(\mathsf{T},\mathsf{V}) = \begin{cases} (x-z) - \alpha(x-w) + \alpha(z-y) & \text{if } w \geq x \\ (x-z) - \beta(w-x) + \alpha(z-y) & \text{if } x \geq w \end{cases}$$

Since $-(x-w) + (z-y) > 0$ by (1)-(3), even though the primary payoff gain $x - z$ is negative, the trustee prefers F in BTGs in which $x > w$ if α is large enough. In this way inequality aversion can explain fulfilment.[7]

In *kindness reciprocity* theories, if E believes R is being intentionally kind to him by his action, this makes E wish to choose an action that is kind to R. How kind R's intention is depends on her belief about what E's choice will be, since this determines R's perception of the effects of her choice on E. Thus kindness-reciprocity theories are psychological game theories.

Rabin's kindness-reciprocity hypothesis is also capable of explaining fulfilment: for some values of the BTG parameters there are possible values of players' beliefs for which kindness-reciprocity implies positive fulfilment. (Rabin himself explains it by showing that there is an equilibrium with a positive fulfilment probability in a psychological game.) Rabin's measure of R's kindness to E (Rabin 1993) is

$$(5) \quad K(t, f^*) = \frac{v(t, f^*) - \overline{v}(f^*)}{v^h(f^*) - v^\ell(f^*)},$$

where $v^h(f^*), v^\ell(f^*)$ and $\overline{v}(f^*)$ are respectively E's highest, lowest (Pareto-optimal) and 'equitable' payoffs given that E plays F with probability f^*, and the 'equitable' payoff is the mean of the first two.[8] If the Mutual Gain condition (4) holds, R is kind for high enough t, and E chooses F provided that the temptation payoff gain $z - x$ is smaller than the secondary utility from reciprocating.[9] The condition is shown in the Appendix to be

$$(6) \quad z - x < (t^* - 0.5)/t^*.$$

Rabin's kindness K is defined in terms of the difference made to E's payoff by R's choice, an intrapersonal difference; in Falk and Fischbacher (1999) the kindness k of R's act depends on an interpersonal difference, the difference between what R expects E to get from that act, and to get from it herself. R's kindness as perceived by E, k^*, is then given by

$$k^* = v(t^*, f^{**}) - u(t^*, f^{**}),$$

where u, v are the primary payoffs of R and E. Once again, kindness reciprocity can explain fulfilling. The act T is perceived kind provided that $v(1, f^{**}) - u(1, f^{**}) > 0$ or

$$(7) \quad f^{**}(x - w) + (1 - f^{**})(z - y) > 0,$$

Condition (7) can easily hold in BTGs, since the definition of a BTG puts no restriction on either $x - w$ or $z - y$. There is no need for the Mutual Gain condition, since a, b do not enter (7).

Since inequality aversion, 'kindness' in two guises, and trust responsiveness can all explain fulfilling, it is important that we should be able to discriminate between them if fulfilling is observed. This discrimination is simplified if not only fulfilling but also trust responsiveness is observed, for trust responsiveness is essentially incompatible with kindness-reciprocity. Since (6) does not involve f^{**}, Rabin's model fails to predict trust-responsiveness, and since (1), (2) and (3) give $x - w < z - y$, (7) implies that Falk-Fischbacher perceived kindness actually decreases with f^{**}.[10]

1.6 Attitudinal Theories

We shall call any theory of trust and fulfilment in which the trustee's choice depends on how favorably he regards the truster's action an *attitudinal* theory. Kindness-reciprocity

theories are attitudinal because in them E's preference between F and V depends on how kind he thinks a T choice is, and he has a pro attitude to kind acts. But kindness is not the only feature of R's choice that might provoke an attitude-driven motive to fulfil in E. For example, E might feel that R had a greater or lesser *need* to depend on him. Compare the thoughts of the peasant trusted not to give away the prisoner-of-war and those of Hausman's (1998) trustee who is requested in a note to feed the cat of a neighbor who has taken off for the weekend on an impulse. These are BTGs in which trusting is Rabin-unkind (inducing a con attitude in E), but in the first it is also needful, inducing a pro attitude which may more than compensate the perceived unkindness.

Cases of the prisoner-of-war kind are characterized, intuitively, by large negative values of the truster's *status quo* payoff a. For this reason we will call the magnitude $-a$ the 'need to trust' of the truster. Of course, even quite a large $-a$ does not guarantee that R will be seen as in need. Moreover, there are other ways in which a negative a could induce a pro attitude to a T choice; for example T might be seen as a justifiable attempt to equalize an unequal distribution, or to maximize the sum of payoffs. It is also possible that high $-a$ might militate *against* fulfilment, e.g. by reducing R's 'exposure' $a - y$. Such a reduced-exposure effect is hypothesized by Rigdon *et al.* (2000) and Pelligra (2000).[11]

Our stance is that one plausible effect of a high $-a$ is that E sees R as having a need to play T. Schotter *et al.* (1996) find that in ultimatum games low offers are more likely to be accepted when offerers are only allowed to participate in the second stage of the experiment if they secure a large share. They suggest that lower proposals are seen by receivers as justified by a 'need to survive'. Such a 'need' is induced by a reference point, in this case 'staying in business', and in the BTG breaking even in the interaction. However, despite the plausibility of a perception of need we intend the label 'need to trust' only as a shorthand: it refers to perceived need together with any other properties of $-a$ which might affect attitudes to T and so the propensity to fulfil.

In Rabin's and Falk and Fischbacher's theories E's choice reflects his attitude in a direct and simple way: the more pro his attitude is the more he prefers an act which benefits R. But pro and con attitudes of E towards R might also affect E's willingness to fulfil in an indirect way; they might interact with his estimate f^{**} of R's confidence. This is because, as we argued in Subsection 1.4, the degree of trust responsiveness is likely to

depend on the sympathy or respect the trustee feels for the truster. Since sympathy and respect are likely to be enhanced both by perceived kindness and perceived need, we might expect a higher degree of trust responsiveness when in games in which trusting is kind or needful.[12] We label the hypothesis that sympathy, respect and other pro attitudes strengthen trust responsiveness the *Interaction Hypothesis*.

It might be conjectured that E's attitude to R's choice should make no difference to E's preference between F and V when $f^{**} = 0$. This attitude independence at $f^{**} = 0$ is predicted by a 'rewarding theory' of fulfilment which says that E's secondary motive for choosing F or V is to reward or sanction R, according to his attitude to her choice. The reasoning is as follows. When $f^{**} = 0$, E will typically be sure that R will choose W. But then, since if R does so her payoff is unaffected by E's choice, E must also think there is no scope for rewarding or sanctioning R by his action. Hence his propensity to fulfil must be determined on other grounds than his attitude. And so, as trustees' attitude to trusting varies with the parameters of the BTG, f should remain unaltered when f^{**} vanishes. Geometrically, the graph of the response of f to f^{**} for BTGs with different payoff characteristics would all have the same vertical intercept; we therefore label this the *Common Intercept Hypothesis*.

The Common Intercept Hypothesis says nothing about the height of the intercept, or even whether it is positive or zero. Clearly, if trust responsiveness were the *only* force at work in trust games, it would be zero. But trustees might choose F for reasons unconnected with f^{**}. They might tremble, or choose F as an 'expressive' act (Hargreaves Heap *et al.* 1992). In these cases f might be positive at $f^{**} = 0$. We call this the *Positive Intercept Hypothesis*. Some findings in the literature, which suggest that there is a type of player whose tendency to fulfil trust is rather rigid (e.g. Glaeser *et al.* 2000), support the Positive Intercept Hypothesis.

2. DESIGN

2.1 Main Features

In the experiment we tested for trust responsiveness by observing trustees' rates of fulfilling, f, measuring their perceptions f^{**} of the confidence of the truster, and estimating the former as a function of the latter and of other variables. We also sought to determine

whether trust responsiveness is affected by changes in the parameters of the BTG. The design had four salient features: (i) three different versions of BTG were administered; (ii) choices of strategies were elicited; (iii) certain first and second order beliefs about choices were elicited; (iv) each subject in the E role, before being asked to estimate his co-player's confidence, received good quality information relating to it, in the form of a 'report'.[13] We comment on these features in turn.

2.2.1 The three BTG variants. In order to manipulate perceived kindness and perceived need to trust, after piloting we selected three parametrizations of the BTG. These are shown in Table 2. The entries represent gains and losses of money in units of £1.

Table 2: THREE VARIANTS OF THE BASIC TRUST GAME

	Gratuitous(GTG)		Kind(KTG)		Needy(NTG)	
	Trustee (E)		*Trustee (E)*		*Trustee (E)*	
Truster (R)	F	V	F	V	F	V
T	$3,3$	$-3,4.5$	$3,3$	$-3,4.5$	$3,3$	$-3,4.5$
W	$0,3$	$0,3$	$0,0$	$0,0$	$-1.5,0$	$-1.5,0$

In the Kind Trust Game (KTG) choosing T has positive Rabin kindness. The Gratuitous Trust Game (GTG), introduced in Pelligra (2000), has the same parameters as the KTG except that $b = 3$, making T have zero Rabin kindness when R expects F. The Needy Trust Game (NTG) is the same as the KTG except that $a = -1.5$. Since the three BTG variants have identical top rows, the effect of all row-defined psychological motives for fulfilling is constant across variants: in particular, neither inequality aversion nor perceived Falk-Fischbacher kindness could account for any variations we might observe in the rate of fulfilling across variants.

2.2.2 The strategy method. We use the strategy method of Selten (1967): we ask subjects to choose between strategies, which are then played out to determine the outcome. This method requires the E player to think about what he would do *if* R were to choose T. One advantage is that it provides data on E's preferences at unreached nodes. Another arises from the nature of trust responsiveness. To test for this we need to measure E's

confidence-perception, f^{**}, at the time of her decision whether or not to fulfil trust. If she took this decision after observing that R had chosen T, f^{**} would also have to be measured after her observation of T. This would have the effect of truncating the range of variation of the independent variable f^{**}, since most subjects would conclude from seeing T that R's confidence f^* was high. For example, if E thinks, game theory-wise, that R maximizes her expected payment (and is risk-neutral), E should conclude from seeing T that f^* can not be less than the critical value

$$(8) \quad f_{\mathsf{crit}} = \frac{a-y}{w-y},$$

which is equal to 0.25 in NTG and to 0.5 in GTG and KTG. Against these advantages, there is some evidence that second-stage choices in games played out sequentially may differ from the choices implied by the strategies chosen under the strategy method (Schotter *et al.* 1994, Shafir and Tversky 1992); but Brandts and Charness (2000), Cason and Mui (1998) and Güth *et al.* (in press) find no statistically significant difference.

2.2.3 Eliciting beliefs. Like Dufwenberg and Gneezy, we measured R players' beliefs about whether their coplayers would choose to fulfil, and E player's beliefs about these beliefs, by direct elicitation schemes.[13] But the belief variables were quite different from those in their study. In a BTG the fulfilment variable is dichotomous, so R's confidence can be naturally measured by a single number, f^*, R's probability for F. In Dufwenberg and Gneezy's experiment, however, as in the other earlier experiments we have discussed, fulfilment is a many-valued variable, y say, and this measure is not available. Instead, it is natural to define R's belief that her trust will be fulfilled, as they do, by the expectation Ey. Our second order belief variable is, like theirs, the E player's expectation of the number between 0 and 1 which is the outcome of the first-order elicitation (in our case, the co-player's response, in theirs the average of such responses).

An advantage of the BTG is that the confidence variable, f^*, describes R's belief state in an unambiguous way. A given value of the measure Ey, on the other hand, is compatible with many subjective probability distributions over the support of y. This ambiguity infects the corresponding measure of E's belief about R's belief. It is not clear that one should expect the same response to a given value of the second order expectation, whatever distributions lie behind it.

2.2.4 The E player's report. In formulating hypotheses about belief-driven motives

it has been the common practice (e.g. Geanakopolos *et al.* 1989, Rabin 1993, Falk and Fischbacher 1999) to represent beliefs by point estimates, as we too have done. But this suppresses an important aspect of beliefs, the weight of evidence upon which they are based. It is reasonable to suppose that a person will display significant trust responsiveness only when she has definite beliefs, based on evidence of good weight, about R's confidence; and in particular that 'ambiguity' about R's confidence might tend to disable the mechanism. If this is so, there would be little point in testing for trust responsiveness in a setup in which most E players felt that their estimates were mere guesses. We therefore sought to provide E players with evidence. No doubt the best information about the confidence of a particular R player is in the head of that subject, but extracting it without distortion presents problems: if that player knew that her report would be conveyed to her coplayer, she would sometimes have a strategic motive to misrepresent. For example, if she thought her coplayer might be trust responsive, she would have a motive to exaggerate. To deal with this our design uses 'motivated cross-talk': each E subject is informed not of his own coplayer's stated value of her confidence, but of a summary statistic of the stated confidences of other R subjects.[15]

2.3 Structure and Procedure

The experiment was run in the Department of Economics in the University of Oxford in February 2001. Recruiting was by an advertisement saying that participants would be taking part in a scientific experiment on interactive decision making, and would be paid an amount depending partly on their decisions and partly on chance. Recruits were predominantly undergraduate or graduate students, but some were in university or other jobs. There were 10 sessions in the main experiment. Each session involved eight subjects, four in the R role and four in the E role. Subjects responded to computer-administered instructions, the full text of which appears in Appendix 2. Each subject played four rounds of the BTG, one with each of the subjects in the other role.[16] Rounds 1 and 2 were plays of one of the three variants of BTG (GTG, KTG, NTG) and rounds 3 and 4 were plays of a different variant. The order in which a given pair of variants was presented was counterbalanced over sessions. In all there were 16 rounds of KTG, and 12 each of GTG and KTG. Since a round devoted to a given variant contained four plays of that variant, one by each of four pairs of subjects, there were in all 64 plays of

KTG and 48 each of GTG and KTG, and thus 160 values for each behavior variable. Before the experiment subjects supplied demographic details, of age, sex and occupation and, if students, their course. At the end of the session subjects were invited to make written comments.

A session consisted of three stages. At the start of it, subjects were assigned randomly to terminals separated by screens. In the Introduction Stage, the nature of the tasks and the payment procedure were explained, with examples and practice. Next, four subjects were assigned randomly to the R role and four to the E role. The Play Stage now began. In each of the four rounds the order of events was as follows.

1. Each subject was shown the payoff matrix of the variant of BTG to be played, in the form of a 'points table'.

2. Each R player made a *statement*, s, of the probability she attached to the event that her coplayer would choose strategy F.

3. Each E player received a *report*, r, consisting of the mean value of the statements of his non-coplayers.

4. Each E player made a *guess*, g, at the statement of his coplayer.

5. Each player made her BTG strategy choice.

The statement s measures the R player's confidence f^*, and the guess g measures the E player's belief about his coplayer's statement and so measures his confidence-perception f^{**}. Statements and guesses were made by using the mouse to manipulate a pointer on a semicircular dial calibrated in integers from 0 to 100. Reports were rounded to the nearest integer.

Subjects were told nothing at the end of rounds 1, 2 or 3 about the strategy choices of others in either the current or earlier rounds. The only information any subject received about other subjects' behavior was the report about R players' statement given to E players.

In the Payment Stage, two rounds were chosen at random. Subjects were paid for the strategy choices they had made in one of these, in accordance with the points table. For the other randomly chosen round, they were paid for their statements (if R players) or guesses (if E players). Statements were paid according to the well-known quadratic

scoring rule (Davis and Holt 1993), and guesses according to a triangular scheme.[17] Sessions averaged about 55 minutes; subjects' total payments ranged from £1.11 to £10.00 and averaged £5.92.

Strategy payments could be negative. In order to maximize the psychological impact of the negative values while ensuring that nobody left the experiment out of pocket, subjects were given, instead of the usual unlosable turn-up fee, an 'initial credit' (of £4), and told that they might either add to it or lose it during the experiment.

2.4 Hypotheses

Our general purpose is to establish whether there is such a thing as trust responsiveness and, if so, how its strength varies with the payoff parameters of trust games and how it is related to other forces which may motivate the fulfilment of trust. Our earlier discussion of the factors that may work for or against fulfilment raises several specific questions. These can conveniently be expressed in terms of hypotheses about the *fulfilment function*, the function giving the trustee's propensity to fulfil f in terms of factors that varied in the course of the experiment. We call the gradient of the fulfilment function, $\partial f/\partial f^{**}$, the *coefficient of trust responsiveness*. Of particular interest are six hypotheses.

H1 (*Positive Propensity*) The average propensity to fulfil f is positive.

H2 (*Variable Propensity*) The average propensity to fulfil varies with the BTG variant.

H3 (*Trust Responsiveness*) The coefficient of trust responsiveness is positive.

H4 (*Interaction*) The coefficient of trust responsiveness is lowest in GTG, greater in KTG, and greatest in NTG.

H5 (*Variable Intercept*) The propensity to fulfil at $f^{**} = 0$ varies with the BTG variant.

H6 (*Positive Intercept*) The average propensity to fulfil is positive at $f^{**} = 0$.

H7 (*Personal Characteristics*) The propensity to fulfil varies with demographic variables.

H1 expresses the now well corroborated finding discussed in Section 1, contradicting the conventional wisdom that players are rational maximizers of monetary rewards. H2 is implied by both kindness reciprocity and trust responsiveness theories, since they

both make fulfilment depend on the BTG parameters we vary. H2 is denied by a pure inequality aversion theory, in which fulfilment depends only top row payoffs, which we hold constant. H3 is our own central hypothesis. A positive coefficient of trust responsiveness means that there is trust responsiveness, always provided that the association is produced by a causal relationship running from f^{**} to f. H3 is inconsistent with standard kindness-reciprocity theories. H4 is the form that the Interaction Hypothesis naturally takes in the present experiment, since trusting is kind in KTG and NTG but not in GTG, and in NTG the truster has in addition a need to trust. The assumption that adding a 'need to trust' by making a negative increases the pro-ness of E's attitude is false if the reduction in the exposure from trusting has a strong enough negative effect. H5 is the negation of the Common Intercept Hypothesis and H6 is the Positive Intercept Hypothesis. H7 collects several hypotheses, corresponding to the various demographic data variables we collected.

The account of trust and trustworthiness in BTGs which we outlined in Section 1 led us to expect to find support for H1, H2, H3 and H4 and to have open minds about H5, H6 and H7.

3. RESULTS

3.1 Commentary on the Raw Data

Table 3 shows some summary statistics. It is clear that we can reject the analogue of H1, the hypothesis that there is no trust. The overall proportion of T choices was 0.49, and it was significantly positive in each variant of the BTG: the mean rates of trusting in GTGs, KTGs and NTGs are 0.33, 0.52 and 0.61 respectively, which all have $p < 0.001$. Even in the Gratuitous Trust Game, where R cannot believe that E will choose F to reciprocate perceived kindness, nearly one third of choices were trusting choices. We shall see soon how much of this trust was warranted. The differences between the means of statements and reports, theoretically equal, are due to the rounding in the reports.

Table 4 shows the raw data for trusters. A striking feature is the variability both across and within subjects of the statement s, the elicited values of the R players' expression confidence. Writing s_i for the statement of the ith R player, and $\overline{s}_i$ for its within-subject mean, the standard deviations of $\overline{s}_i$ are 0.26, 0.28 and 0.28 in GTG,

Table 3: Means and Standard Deviations of Observed Variables

	GTG	KTG	NTG	All
Trusting	0.33	0.52	0.61	0.49
	(0.48)	(0.50)	(0.49)	(0.50)
Statement	0.28	0.38	0.32	0.33
	(0.28)	(0.32)	(0.32)	(0.31)
Report	0.27	0.39	0.34	0.34
	(0.18)	(0.18)	(0.23)	(0.20)
Guess	0.29	0.43	0.39	0.38
	(0.24)	(0.25)	(0.30)	(0.27)
Fulfilling	0.27	0.40	0.52	0.40
	(0.45)	(0.50)	(0.50)	(0.49)

KTG and NTG respectively, and even the within-subject standard deviations of s_i have means of 0.10, 0.13 and 0.13. One explanation of the inter-subject variance is that subjects came to the laboratory with widely dispersed views of human nature. We remark that the empirical distribution of subjects' beliefs about others' choices in the BTG looks very unlike an equilibrium of a psychological game. In models of this kind (Bacharach and Gambetta 2001b, Dufwenberg in press), any equilibrium features a common distribution over choices and a common point estimate f^*. The within-subject variations often appear to be without obvious sense – as McKelvey and Palfrey (1992) remark, "there are interesting *non-patterns* in the data ..., common irregularities ... which appear rather haphazard". There are often big jumps between subjects' two s values for a given variant, even though the subject has received no new information. These jumps may give the impression that statements were often based on mere whim, or that subjects were revising their estimates as new considerations occurred to them. Other cases involve apparently perverse changes of choice, switches from W to T going with large reductions in statements of confidence, or *vice versa*.[18]

However, there are also strong patterns in the data. Although the classic prediction of game theory is falsified by our data, the more nuanced prediction of game theory that R players are fallible maximizers of expected payoff given their beliefs about coplayers' choices is not inconsistent with the data. One can explain the substantial trust rates as the effect of R players being faithful, if noisy, subjects of game theory who believe that E players are quite likely not. Higher values of stated confidence s are associated

Table 4: RAW DATA: *R* PLAYERS

Ses[a]	*Sub*[b]	*Variant*				*Statement*[c]				*Choice*			
		round				*round*				*round*			
		1	2	3	4	1	2	3	4	1	2	3	4
1	R1	K	K	G	G	25	0	25	0	W	W	W	W
	R2	K	K	G	G	0	0	0	0	W	W	W	W
	R3	K	K	G	G	20	25	15	15	T	T	W	W
	R4	K	K	G	G	80	90	10	10	T	T	W	W
2	R5	K	K	N	N	5	5	5	5	W	W	W	T
	R6	K	K	N	N	13	95	2	3	T	W	T	W
	R7	K	K	N	N	25	25	30	10	T	W	T	T
	R8	K	K	N	N	0	50	0	90	W	W	T	W
3	R9	G	G	N	N	50	50	50	50	T	T	T	T
	R10	G	G	N	N	50	0	50	50	W	W	W	W
	R11	G	G	N	N	25	25	98	1	T	T	W	T
	R12	G	G	N	N	75	75	25	26	T	T	T	T
4	R13	N	N	G	G	0	0	0	0	W	T	W	W
	R14	N	N	G	G	10	0	35	25	W	W	W	W
	R15	N	N	G	G	20	30	25	25	T	T	W	T
	R16	N	N	G	G	0	10	10	5	W	W	W	W
5	R17	G	G	K	K	0	10	35	25	W	W	W	W
	R18	G	G	K	K	50	96	100	100	T	T	T	T
	R19	G	G	K	K	95	75	10	85	T	T	W	W
	R20	G	G	K	K	75	4	7	22	T	T	T	T
6	R21	N	N	K	K	15	35	65	55	T	T	T	T
	R22	N	N	K	K	20	10	12	8	W	W	W	W
	R23	N	N	K	K	25	10	90	50	T	W	T	T
	R24	N	N	K	K	1	20	75	60	T	T	T	T
7	R25	G	G	K	K	36	67	70	62	T	W	T	W
	R26	G	G	K	K	25	0	25	35	W	W	W	W
	R27	G	G	K	K	40	70	20	20	T	T	T	T
	R28	G	G	K	K	0	0	0	0	W	W	W	W
8	R29	N	N	K	K	50	25	50	29	T	T	W	T
	R30	N	N	K	K	90	90	80	40	W	T	T	T
	R31	N	N	K	K	33	25	15	33	T	W	W	T
	R32	N	N	K	K	65	65	65	65	T	T	T	T
9	R33	K	K	G	G	33	75	20	33	W	T	W	W
	R34	K	K	G	G	0	0	25	25	W	W	W	W
	R35	K	K	G	G	10	10	5	5	W	W	W	W
	R36	K	K	G	G	25	20	5	5	T	T	W	W
10[d]	R38	K	K	N	N	0	88	80	100	T	T	T	T
	R39	K	K	N	N	20	20	0	0	W	W	W	W
	R40	K	K	N	N	75	90	100	50	T	T	T	T

[a]Session number [b]Subject code [c]Expressed as a percentage
[d] R37's computer failed.

with more frequent T choices; the mean s values of T-choosers and W-choosers are 0.47 and 0.20 respectively. On the hypothesis that R maximizes expected money payoff and so chooses T only if $s \geq f_{\mathsf{crit}}$ and W only if $s \leq f_{\mathsf{crit}}$, the proportions of wrong T choices were 0.38 in GTG, 0.44 in KTG and 0.36 in NTG, and those of wrong W choices were 0.04, 0.17 and 0.33 respectively. If one assumes a modicum of risk-loving these rates are consistent with expected utility maximization and an error rate of the order of magnitude reported in other studies.[19]

Table 5 shows the raw data for E players. The *report* to an E player in a round is the mean of the confidence statements of his non-coplayers in that round; his *guess* is the elicited value of his estimate of his current coplayer's stated confidence.

It comes as no surprise that H1 is strongly falsified. The overall rate of fulfilling, $\overline{f}$ is 0.40, with means for GTG, KTG and NTG of 0.27, 0.40 and 0.52 respectively. These rates are quite high, remembering that fractional fulfilment is not possible.[20]

We can now see whether trusters were on average overconfident or underconfident, or had correct expectations. The mean statements in GTG, KTG and NTG were 0.28, 0.38 and 0.32 respectively. Thus the average R player got the F rate about right in GTG and KTG, but somehow managed to underestimate the F rate in NTG by quite a wide margin; the greater tendency to fulfil in the NTG than in the KTG that *we* correctly conjectured was for some reason lost on *the player*. One possibility is that an R player type with an egoistic social value orientation, and so unimpressed by need, is more prone than other types to assume that others are like themselves, as the triangle hypothesis (Kelly and Stahelski 1970) has it. The mean values of the guess g in GTG, KTG and NTG were 0.29, 0.43 and 0.39 respectively. The combination in NTG of modest values of g with high values of f, is a first indication that f^{**} may be a more powerful force for fulfilling in NTG than in other variants.

Next we remark that the variability of guesses g is lower, both between and within subjects, than that of statements s. One possible explanation is that E subjects, though starting from the same prior beliefs about fulfilling propensities as R subjects, were drawn towards the estimates of these propensities conveyed in the much less variable reports. Here is a first hint from the data that E subjects took their reports seriously. It can also be seen from Table 5 that a number of many subjects' guesses closely tracked the report. The Pearson correlation coefficient of r and g across all tasks is 0.47. In

Table 5: Raw Data: E Players

Ses	*Sub*	*Variant*				*Report*[a]				*Guess*[a]				*Choice*			
		round				*round*				*round*				*round*			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	E1	K	K	G	G	33	8	11	8	50	37	16	58	F	V	V	V
	E2	K	K	G	G	41	38	13	3	35	30	10	0	V	V	V	V
	E3	K	K	G	G	35	38	8	5	30	40	10	100	F	V	V	V
	E4	K	K	G	G	15	30	16	8	10	5	10	40	V	V	V	V
2	E5	K	K	N	N	12	41	2	35	20	50	20	50	F	F	F	F
	E6	K	K	N	N	10	56	12	32	80	99	50	75	F	F	F	F
	E7	K	K	N	N	6	26	10	4	10	30	13	8	V	F	F	F
	E8	K	K	N	N	14	50	11	34	10	20	10	34	V	V	V	V
3	E9	G	G	N	N	50	25	41	25	50	25	33	25	V	V	V	V
	E10	G	G	N	N	50	33	66	42	0	0	0	0	V	V	V	V
	E11	G	G	N	N	58	50	57	33	35	43	50	40	V	V	F	F
	E12	G	G	N	N	41	41	57	25	5	15	50	30	V	V	F	V
4	E13	N	N	G	G	10	10	15	10	20	5	15	13	F	F	F	F
	E14	N	N	G	G	6	13	20	10	100	100	80	70	F	F	F	F
	E15	N	N	G	G	3	13	23	16	5	11	23	20	V	V	V	V
	E16	N	N	G	G	10	3	11	18	0	0	0	10	V	V	V	V
5	E17	G	G	K	K	74	60	47	44	26	70	47	4	V	V	V	V
	E18	G	G	K	K	58	58	48	49	60	50	50	50	V	V	F	V
	E19	G	G	K	K	41	29	39	70	36	34	35	64	F	F	F	V
	E20	G	G	K	K	49	36	17	69	73	27	75	70	F	F	F	F
6	E21	N	N	K	K	15	18	50	55	10	20	33	50	V	V	V	V
	E22	N	N	K	K	13	13	55	41	38	38	46	38	F	F	V	V
	E23	N	N	K	K	12	21	59	37	10	25	40	40	V	V	V	V
	E24	N	N	K	K	20	21	76	39	20	20	70	75	V	F	F	F
7	E25	G	G	K	K	21	45	31	27	50	50	25	25	V	V	V	V
	E26	G	G	K	K	25	23	38	32	21	18	31	26	V	V	V	V
	E27	G	G	K	K	20	45	15	39	23	15	24	15	F	V	F	F
	E28	G	G	K	K	33	22	30	18	0	0	0	0	V	V	V	V
8	E29	N	N	K	K	62	46	65	42	75	25	65	65	V	V	V	V
	E30	N	N	K	K	49	60	48	44	25	60	50	50	V	F	V	V
	E31	N	N	K	K	68	38	53	34	93	97	95	96	F	F	F	F
	E32	N	N	K	K	57	60	43	46	66	70	32	35	V	V	V	V
9	E33	K	K	G	G	11	28	16	14	30	30	30	25	F	F	F	F
	E34	K	K	G	G	22	10	16	21	40	40	35	35	F	F	F	F
	E35	K	K	G	G	19	35	11	21	15	15	10	10	F	F	F	F
	E36	K	K	G	G	14	31	10	11	20	50	25	20	V	F	V	V
10	E37	K	K	N	N	31	66	93	33	30	80	90	75	V	F	F	F
	E38	K	K	N	N	56	66	60	66	55	66	60	66	F	F	F	F
	E39	K	K	N	N	50	66	60	50	75	68	70	33	F	V	F	F
	E40	K	K	N	N	31	89	66	50	25	100	25	35	V	V	V	V

[a] Expressed as a percentage

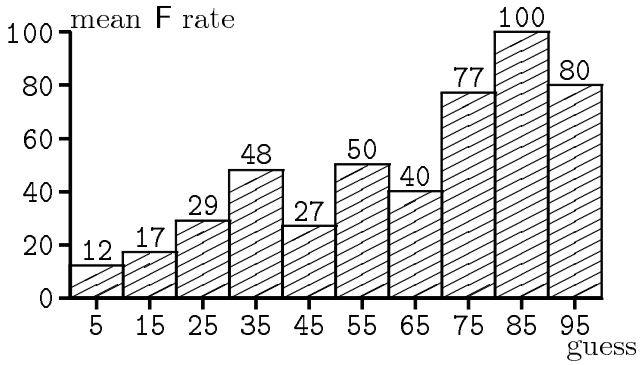
Subsection 4.2 we shall look in more the detail at how guesses were influenced by the incoming report.

Were subjects trust responsive? We can now get a first inkling of the answer the experiment provides to this, our central question. A rough and ready measure of overall trust responsiveness is *crude trust responsiveness* $\mathsf{ctr} = (\overline{g}_{\mathsf{F}} - \overline{g}_{\mathsf{V}})/\overline{g}_{\mathsf{V}}$ where $\overline{g}_{\mathsf{F}}$ (resp. m_{V}) is the mean value of g in the subset of subjects who chose F (resp. V).[21] The magnitude ctr equals 0.68 over the whole sample, *prima facie* evidence in favor of H3. Its values are 0.47, 0.44, and 1.19 in variants GTG, KTG and NTG respectively, so there is mixed evidence concerning H4, the Interaction Hypothesis.

Figure 1 shows the relative frequency of F choices (in all variants combined) in different intervals of g values. We see that this relative frequency rises more or less monotonically from about 0.2 to 0.8, suggesting a sizeable coefficient of trust responsiveness, of the order of 0.6, and some support for the Positive Intercept Hypothesis H6. But inferences from Figure 1 must be tempered with caution.[22]

Neither the ctr rates nor Figure 1 make any attempt to separate out between-subject and within-subject variation. Trust responsiveness is an intrapersonal variation, so to establish it one needs to make sure that an observed response of gross f to gross g is not an artefact of interpersonal differences that happen to be correlated with both f and g. The next subsection presents a more refined analysis, which resolves this and other questions not answered by our preliminary inspection.

Figure 1: Variation of Fulfilling Rate with Guess



3.2 Econometric Analysis

In the last subsection we drew some tentative conclusions from the raw data by piecemeal application of naive statistical measures. These measures decisively rejected H1, and appeared to provide some support for H3, H4 and H6.

Table 6: Fulfilment Function: Estimate of Catholic Model

Variable	*Coefficient*	*S.e.*	*p*
const	−0.8423	0.6810	0.216
g_G [a]	0.0177	0.0085	0.038
g_K [a]	0.0220	0.0082	0.008
g_N [a]	0.0267	0.0092	0.004
sus [b]	0.8380	0.4242	0.048
GTG [c]	−0.2532	0.4792	0.597
KTG [c]	−0.3261	0.5344	0.542
male [d]	−0.1456	0.3577	0.684
dage [e]	0.0985	0.0561	0.079
grad [f]	−0.5473	0.4255	0.198
hum [f]	0.1700	0.7264	0.815
sci [f]	0.4925	0.7136	0.490
socsci [f]	0.0881	0.6781	0.897

[a] $g_v = g$ in vTG games, otherwise 0 (v = G,K,N).
[b] Dummy for suspect subject.
[c] vTG = 1 in vTG games, otherwise 0 (v = G,K,N).
[d] Dummy for male subjects (49 percent).
[e] Age minus mean subject age. Ages ranged from 18 to 46, with mean 24.
[f] Dummies for graduate, humanities, science and social science students. 86 percent of subjects were students, of whom 42, 27, 34 and 39 percent were of these categories.

In this subsection we report an econometric analysis in which we simultaneously estimated the dependence of the propensity to fulfil f on a broad range of variables. We estimated a probit model of the form

$$f = \Phi(\boldsymbol{\beta}'\boldsymbol{x})$$

where $\Phi(y)$ denotes the probability that a standard normal variate is less than y. We began by estimating a catholic model in which the vector $\boldsymbol{x}$ included a broad range of the explanatory variables on which we had data and which are suggested by past

experiments, theory and our own conjectures. We iteratively eliminated the explanatory variables that failed to pass a significance test, using a significance level of 0.05, and reintroducing eliminated variables to test for revivals of significance. Table 6 shows the result of estimating the catholic model.[23] The explanatory variables include dummies for game variants, demographic variables, and the E player's guess g, which acts as a proxy for f^{**}. The latter is differentiated by the variant being played, to allow interaction effects to be picked up if there are any. The variable g_v is defined to take the value of g in variant v ($v = G, K, N$, in obvious notation), and zero in other variants. The Interaction Hypothesis says that $g_G < g_K < g_N$.

In most cases the signs make sense, but only four explanatory variables are significant at 0.05: g_G, g_K, g_N and *sus*. The variable *sus* is a dummy for a few anomalous subjects who we suspected might choose F or W for reasons quite extraneous to the theory.[24] The most promising demographic variable is *dage*, the subject's age as a deviation from the mean over subjects, which is significant at 10 percent.[25] The dummies for the variants are notable for their lack of significance.

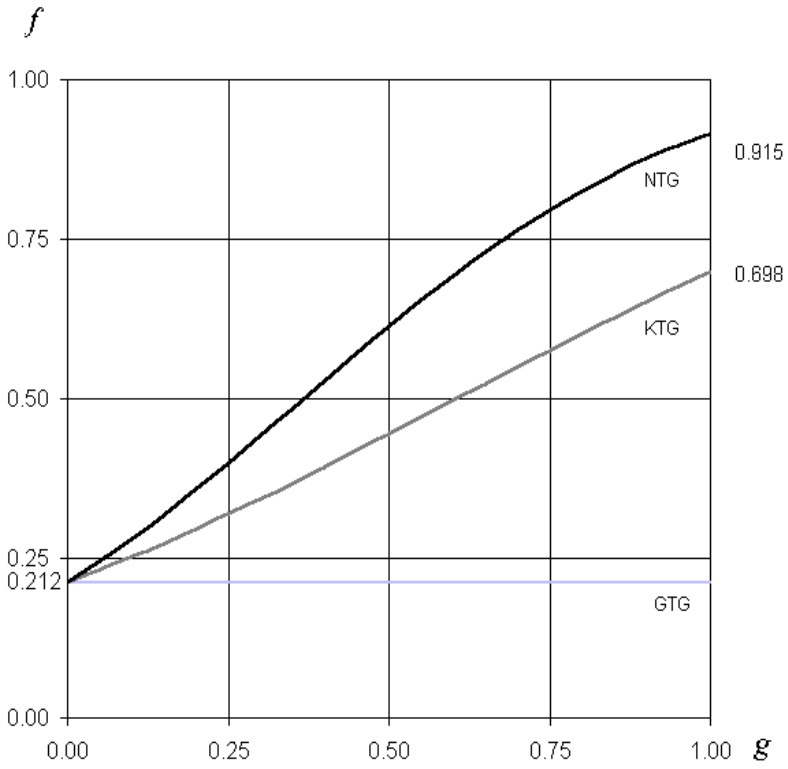
The broad pattern of explanation that emerges from the catholic regression persisted with only minor variations throughout the elimination process. The only statistically significant effects in the data proved to be interaction effects between the game variant and the E player's guess, and a positive intercept. The estimated final regression is

$$(9) \quad f = \Phi(-0.80 + 1.32g_K + 2.17g_N + 0.91sus)$$

Figure 2 shows (9) for normal subjects graphically. We see that in two out of three variant treatments subjects are trust responsive, with trust responsiveness lowest (zero) in GTG, greater in KTG, and greatest in NTG (f rises by 0.49 in KTG and by 0.70 in NTG as g goes from 0 to 1). There is positive intercept of 0.21 which is the same for all three variants. Figure 2 thus strikingly bears out hypotheses H1 (Positive Propensity), H2 (Variable Propensity), H3 (Trust Responsiveness), H4 (Interaction) and H6 (Positive Intercept) and denies H5 (confirms Common Intercept). The absence of further variables from (9) denies the hypothesis H7 that demographic variables matter. The estimated rates of trust responsiveness in the two variants with kindness are very substantial.

Table 7 displays the coefficients of (9) together with their standard errors and p values, and the estimated coefficients of trust responsiveness $\partial f/\partial g$ at $g = 0.5$. The

Figure 2: Fulfilment as a Function of Guess in the Three Variants



probit form means that the coefficients in (9) cannot be identified with the gradients of the dependent variable with respect to the regressors. Equation (9) says, for example, that if a normal subject's guess goes up from 20 to 30 in a KTG game, this raises the probability of F from $\Phi(-0.8004 + 1.32 \times 0.20)$ to $\Phi(-0.8004 + 1.32 \times 0.30)$, that is, from 0.296 to 0.343 – a gradient $\partial f/\partial g$ of roughly 0.47.

In estimating the standard errors of the probit coefficients it is important to allow for the panel nature of our data: each subject provides four observations and the four are unlikely to be independent, but may contain powerful 'individual effects'. This expectation is confirmed by an analysis of variance. We find that the indicator variable for fulfilling (1 for an F choice and 0 for a V choice) has between-subject and within-subject sample variances of 0.703 and 0.092; the hypothesis of equal variance yields $F(39, 120) = 7.664$ and $p < 0.001$. The estimation method must take this into account, or else the data set is treated as if it contained more independent information than

it really does, and p values are underestimated. The standard errors here are therefore 'robust' standard errors based on a 'sandwich' estimator that allows for arbitrary correlations among observations on a single individual (Guilkey and Murphy 1993).

[

Table 7: FULFILMENT FUNCTION: ESTIMATE OF FINAL MODEL

Variable	*Coefficient*	*S.e.*	p	$f'(0.5)$
const	-.8004	0.2497	0.001	–
g_{K} [a]	0.0132	0.0049	0 .007	0.522
g_{N} [a]	0.0217	0.0070	0.002	0.831
sus [b]	0.9089	0.3430	0.008	0.909

[a] $g_v = g$ in vTG games, otherwise 0 (v = G,K,N).
[b] Dummy for suspect subject.

An important question is whether the separation of the three curves in Figure 2 is due to real differences in the response of f to f^{**} in the three variants, or merely to chance. The hypothesis that the coefficients on g for NTG and KTG are equal, against the alternate that the former exceeds the latter, although not rejected using our criterion of 5 percent significance, has a p value of only 0.052. At a slightly less strict significance level the same iterative estimation procedure yields the same final equation and formally rejects equality of g_K and g_N. Moreover, the iterative estimation method, although it does not show that eliminated variables have no effect, requires that we maintain the null hypothesis that they have no effect, and so in the present case that g has no effect in GTG. We conclude that the analysis should be interpreted as showing that the effect of g is null in GTG, positive in KTG, and highest in NTG.

The econometric analysis confirms and extends the results of the naive data analysis. H1 is confirmed: there is a positive rate fulfilling in all treatments and for all f^{**}. H2 is confirmed: at all $f^{**} > 0$ the average propensity to fulfil increases from GTG to KTG to NTG. H3 (Trust Responsiveness) is confirmed by the significant positive coefficients on f^{**} in KTG and NTG. Beyond this, the econometrics gives clear support to H6 (Positive Intercept) and support to H4 (Interaction). It also rejects H5 (Variable Intercept) and H7 (Personality), since the regressors for game variants and personal characteristics were all eliminated as insignificant in the iterative estimation process.

4. DISCUSSION

A number of questions are left unanswered by the statistical analysis of observational data that we have just given. We briefly discuss these in this section.

4.1 Interpretation of the Estimated Model

4.1.1 Is (9) too parsimonious? Formal hypothesis tests can deliver misleading answers. In particular, in experimental studies of the present sort there is always a serious risk of type 2 error. Before concluding that, for example, personal characteristics have no effect on the fulfilling rate, it should be remembered both that commonsense tells us otherwise and that, if several factors are in fact at work, some are likely to be eliminated as insignificant because of the limited statistical power of a single experiment of moderate size. In the case of 'main effects' of the variants things are a little different since, as we have suggested, there is an *a priori* reason to expect a common intercept. As for f^{**} in the GTG, prior considerations point to a lower effect than in other variants, and perhaps to a small or even negative one (trustees might actively resent a truster's confidence).

4.1.2 Causality. Trust responsiveness is a causal hypothesis in which it is f^{**} that causes f. We have argued that there are plausible causal mechanisms with this directionality, in particular those postulated by letting-down aversion theories. The econometric analysis of Section 3 lends further support. First, it establishes a necessary condition for trust responsiveness, that f is positively associated with f^{**}. But also, by delivering positive slope for some variants and different slopes for different variants, it supports the causal direction hypothesis against the two most obvious alternatives.

The first of these alternatives is that f and f^{**} are jointly caused by a social norm or collection of norms; in a given context these norms motivate a trustee E to a certain degree towards F, but also, since the force of such norms is common knowledge, E knows that his coplayer knows that he is subject to them. The second alternative is that f lifts f^{**} through a 'double projection' mechanism: a trustee whose personal characteristics dispose him to a certain degree to choose F not only projects this disposition onto his coplayer (Orbell and Dawes 1991), but also assumes that his coplayer projects her disposition onto her coplayer.

The social norm theory predicts, in the case of a single norm for all BTGs, no relationship between f and g; if there is a different norm for each variant, a single point

in (g, f) space for each variant or, allowing for noise, an undirected scatter of points for each variant. What it is unable to explain is an upward slope in the fulfilment function for a given variant. Whatever explains a deviation δf from the norm by a trustee E, there is nothing in the social norm theory which implies that E should expect her coplayer R to expect a deviation correlated with δf, since this deviation is not generated by the norm. To be sure, E might expect R to expect a correlated deviation, but this can only be explained by a different theory, such as a projection theory. The social norm theory is inconsistent with the upward slopes of the fulfilment function in KTG and NTG.

The double projection hypothesis predicts that any given value of f induces a value of g independent of the variant being played or, allowing for noise, a conditional distribution of g given f independent of variant, for all f. But this in turn implies a single conditional distribution of f given g for all g, and so a single curve for all three variants. That is, the double projection theory is inconsistent with the separation of the three variant fulfilment functions. By indicating separation, therefore, our data also render the double projection hypothesis implausible.

4.1.3 Kindness and other influences on attitudes. We have drawn contrasts between the trust responsiveness theory and well-known theories according to which kindness is the key to understanding trustworthiness. But kindness matters in the theory we are advancing. It is central to our findings that Rabin 'kindness' raises trustworthiness, but that it does so only in the presence of perceived confidence. We have also shown that other perceived attributes of the choice of T which depend on the parameters of the trust game matter, and in particular that what we have termed 'need' does. Care is needed in interpreting both these findings. As we have explained, we regard 'need' only as a convenient label for a feature of the payoff structure which also admits of other descriptions. Just the same is true of 'kindness'.[26]

4.2 The Formation and Expression of Beliefs

For our estimated relationship (9) to establish trust responsiveness and related hypotheses such as Interaction, the guess g must be a good proxy for f^{**}. One obvious requirement for this is that when they gave g values E subjects were accurately reporting their beliefs about their coplayers' confidence statements s. They were rewarded for the accuracy of their g values, and had no clear countermotive to misrepresent, so it is

reasonable to suppose that most subjects were doing so honestly,[27] even if some may have done so only approximately.[28]

But there is a further requirement. If the trust responsiveness hypothesis relates f to firmly held rather than vague beliefs about f^*, g is a good proxy only if it expresses a firmly held belief, which requires that E subjects think they have good evidence about f^*. We meant the report r to be so regarded by E subjects.

One condition for this is that E subjects should have faith in the statements s. Although the quadratic elicitation scheme gave R players an incentive to be truthful in their statements, might they also have had a motive to misrepresent for the sake of the game payoff, for example overstating their confidence in order to induce F through trust responsiveness? Three things militate against this: first, the 'cross-talk' feature, which means that an R player's statement has no effect on the report her coplayer receives in the current round; second,[29] the sophistication needed: the R players would have to theorize that high statements from all of them might activate trust responsiveness; and third, the coordination problem in realizing such a joint strategy.

There is reason to believe that E subjects regarded the report r as good evidence. The simplest measure of the influence of r on g is correlation. The coefficient ρ is 0.47, which has $p < 0.001$.[30] A more sophisticated test is whether their beliefs f^{**} were appropriately affected by r or, in view of the above, whether g was. One feature that we ought to find on the hypothesis that E takes r to be evidence is that the guess g should vary across subjects less than the statement s, and we do (Subsection 3.2). Another that we ought to find is that g rises if it was below r in the first round of a variant and r rises in the second round (and g should fall in the symmetrically opposite case). This test applies even to subjects who may have had strong prior views and, for this reason, guesses only weakly correlated with their reports. We find that of the 33 movements in g that occurred in such cases 25 were in the predicted direction, which in a one-tailed test against the null that shifts up and down are equally likely has a p value of 0.002.

5. CONCLUDING REMARKS

5.1 Summary

The 'self-fulfilling property of trust' or 'trust responsiveness' is the tendency for trustees to fulfil trust because they believe they are trusted. In this paper we have described an experiment to test whether trust responsiveness exists. We first examined the interrelations between this and other hypotheses about what may motivate trust fulfilment, including kindness reciprocity and inequality aversion theories. We interpreted the trust responsiveness theory as an 'attitudinal' theory – one in which a player can be motivated by the pro or con attitude he has to the conjectured action of his coplayer. We suggested that trust responsiveness depends on a pro attitude to trusting, and that a pro attitude may be produced by the perceived kindness of trusting and by the perceived need to trust. In the experiment we observed behavior in three different variants of a basic trust game, and we elicited measures of the truster's confidence f^* (her probability for fulfilment), and the trustee's confidence-perception f^{**} (her estimate of f^*). We used 'motivated cross-talk' (reporting to trustees information about the confidence statements of *non*-coplayers) so that the second-order belief f^{**} would be formed on the basis of relevant and credible evidence. In one of the variants (GTG) trust was neither 'kind' nor 'needy', in the second (KTG) it was kind, and in the third (NTG) both kind and needy.

The attitudinal approach to trust games leads to two predictions. The first is that, if trustees are trying to reward or punish their coplayers, the propensity to fulfil will be the same in all three variants when $f^{**} = 0$ (the Common Intercept Hypothesis). The second is that on the usual accounts of what might produce trust responsiveness – the wish not disappoint outcome or person expectations – the effect will be found more strongly in games with a pro attitude to trusting, and hence in those with kindness or need (the Interaction Hypothesis).

Our analysis of the data indicated that (i) trust responsiveness exists, and the coefficient of trust responsiveness (the gradient of f with respect to f^{**}) may be as high as 0.8 in some trust games; (ii) the Common Intercept Hypothesis is true; (iii) the Interaction Hypothesis is true (the coefficients of trust responsiveness are roughly 0.8 in the NTG, 0.5 in the KTG, and zero in the GTG). We found too that the common intercept is positive – there is in the population some propensity to fulfil trust (roughly

0.2) even when it is thought that the truster has no expectation of fulfilment; and that a converse of the Interaction Hypothesis is true – not only is there trust responsiveness when there are kindness and need, but also, without them there is none. On the other hand, it follows from a common intercept that kindness and need are inert *unless* they are accompanied by perceived confidence.

5.2 Implications

Our study has implications for trust theory, experimental game theory, and social and economic policy. Since 'the self-fulfilling property' is an effect on a player's preference of his beliefs about a coplayer's beliefs about his action, demonstrating it shows that the games people play include psychological games. The efficacy of our methods in yielding results in the case of trust games suggests they may be profitably applied to other games which might be of this class, such as bargaining games and social dilemmas. These methods enabled us not only to show existence but also to estimate the quantitative effects on choice of the belief dependence of preferences. On the other hand there are several ways in which our approach could be further refined, for example by attending to the influence of personal characteristics.

Our analysis demonstrates that the fulfilling rate is strongly sensitive to features of the payoff structure which we would expect to provoke pro or con attitudes to trusting in the mind of the trustee. It therefore supports the view that attitudes are important in explaining strategies, advanced by Rabin and others with respect to particular attitudes. However, a given payoff feature of an act can easily give rise to different perceptions of that act and hence to different attitudes to it, with different effects on choice: careful exploration may be needed to disambiguate them.

By showing trust responsiveness, our study shows that there is a potential in several domains, including e-commerce and work payment schemes, for enhancing fulfilment rates – and so in turn warranted trust levels – by facilitating the *transmission of credible signals of trusters' confidence.* A cheap-talk version of the strategy of confidence-signalling is commonplace: "We're counting on you — Please fill in your census form on Tuesday 7th August, 2001"[31]. Trustees appear even to believe that there is scope for enhancing the T-pair by signalling to the truster that they are aware of the truster's confidence, for example by asserting "We know you're trusting us."[32] When the truster

can choose not to play the trust game but instead take some outside option, choosing to play the game may itself provide a credible signal that she expects fulfilment. This 'forward induction' basis for a trustee to infer confidence gives a theoretical explanation of the success of work payment schemes based on trust rather than monitoring, once we add *trust responsiveness* into the equation.[33] More generally, trust responsiveness may lie behind the imperfectly understood phenomenon of 'motivation crowding out' (Frey and Oberholzer-Gee 1997): by introducing financial incentives the principal credibly signals her low confidence that the agent would exhibit prosocial behavior, and the agent slides down the fulfilment function.

APPENDIX 1. Proof of (6)

The Rabin kindness of t given f^* is, from (5),

$$K(t, f^*) = \frac{v(t, f^*) - 0.5[v^h(f^*) + v^l(f^*)]}{v^h(f^*) - v^l(f^*)}.$$

Since (4) holds, F in response to T makes the trustee better off than W and, for every f^*, $v^h(f^*) = \mathsf{T}, v^l(f^*) = \mathsf{W}$. Since $v(t, f^*) = tv(\mathsf{T}, f^*) + (1-t)v(\mathsf{W}, f^*)$,

$$K(t, f^*) = \frac{(t - 0.5)v(\mathsf{T}, f^*) + (1 - t - 0.5)v(\mathsf{W}, f^*)}{v(\mathsf{T}, f^*) - v(\mathsf{W}, f^*)} = t - 0.5.$$

Similarly, since for each t^* R's payoff is maximized by F, $u^h(t^*) = u(t^*, \mathsf{F})$ and $u^l(t^*) = u(t^*, \mathsf{V})$, whence $L(t^*, f) = f - 0.5$.

In Rabin's model, if E is a kindness-reciprocator his all-in payoff is

(10) $V = v + K^*(1 + L)$,

where K^* denotes E's estimate of R's kindness to him, and L denotes E's kindness to R. It is natural to assume that K^* is given as $K^* = K(t^*, f^{**})$. Then E chooses F only if his secondary utility is positive, which requires $t^* > 0.5$, and in this case if and only if $V(\mathsf{F}) > V(\mathsf{V})$, that is, from (10),

$$t^*x + (1 - t^*)b + 1.5(t^* - 0.5) > t^*z + (1 - t^*)b + 0.5(t^* - 0.5),$$

or $t^* - 0.5 > (z - x)t^*$. □

APPENDIX 2. Experimental Instructions

The instructions are divided into three parts (Introduction, Play and Payment). Subjects in the R and E roles receive different instructions.

I. INTRODUCTION [R and E subjects]

Welcome to the Department of Economics. You are about to take part in an experimental study of decision making. You are not allowed to speak to other participants or communicate in any other way. If you want to ask a question, please put up your hand. The experiment is in three stages, called the INTRODUCTION, STAGE 1 and STAGE 2. In Stages 1 and 2 you will be asked to make some decisions. In the Introduction stage, we will explain the nature of the decisions and give you practice. As of this moment, you have been credited with the sum of £4, your STARTING CREDIT. During the course of the experiment you may, through the decisions you make, either add to this, leave it unchanged, or lose some of it. The net amount to your credit at the end of the experiment will be paid to you before you leave the building. In the Introduction stage, we will explain the nature of the decisions and give you practice. Please wait for other participants.

Introduction

There are three kinds of decisions in all in this experiment: INTERACTIVE DECISIONS, LIKELIHOOD DECISIONS and GUESSING DECISIONS. We will call them I DECISIONS, L DECISIONS and G DECISIONS for short. Everyone will make some I DECISIONS, and each of you will make either some L DECISIONS or some G DECISIONS. None of your decisions will be revealed by us to other participants either during the experiment or after it. In your I Decisions (I for Interactive) you will be paired with another participant, your OPPOSITE NUMBER in that decision. In each I Decision you will have a different Opposite Number. You will not be told who your Opposite Number is, either during the I Decision or later. Nor will your Opposite Number know that it is you that she/he is interacting with.

We'll now explain the decisions you will make in more detail, give you some practice on them, and ask some questions to make sure everything is clear to you. We begin

with I Decisions. At any stage of the explanation, and throughout the experiment, you will be able to return to a HELP page which summarizes what you have been told so far, by clicking the button marked HELP.

I Decisions

In an I Decision, you choose between two OPTIONS, and your Opposite Number also chooses between two Options. Before making your choice you see a POINT TABLE like the one [below]. In it, we have labelled your Options TOP and BOTTOM, and your Opposite Number's LEFT and RIGHT.

You	*Opposite number*	
	LEFT	RIGHT
TOP	*1.5*, 0.5	*1.0*, 2.0
BOTTOM	*0*, 2.5	*2.5*, 2.5

The numbers in the table show the number of POINTS you and your Opposite Number would get, for various combinations of choices by you and her/him; the one on the left of the comma, in blue, shows YOUR payoff, and the one on the right of the comma, in grey, shows your Opposite Number's. In this example, if the Option you chose was BOTTOM and the one your Opposite Number chose was RIGHT, you would get 1.5 Points and she/he would get 0.5 Points. Notice that what you get depends on what you both choose. BOTTOM gets you more points if your Opposite Number chooses RIGHT, but fewer if your Opposite Number chooses LEFT. At the end of the experiment participants will be paid in money for one of the I Decisions they make in Stages 1 and 2. The more points a participant scores the more she/he will be paid. We will explain how the payment is determined in more detail a little later.

We'll now ask you a few questions to make sure you are clear about what I Decisions are and what Point Tables show. After you have entered your answer, you will be given the correct answer and an explanation. If you need further help, please raise your hand. [Subjects are asked the following questions: if their answer is wrong, they are told: "FALSE! The correct answer is " and the corresponding answer. If they are right they are told: "TRUE! " and the corresponding answer].

Q.1 What would you get if you chose TOP and your Opposite Number chose LEFT? Enter a number in the box. [*Answer.* 1.5, because this is the BLUE number in the TOP row and the LEFT column.]

Q.2 What is the worst possible outcome for you in terms of Points if you chose BOTTOM? [*Answer.* 0. If you chose BOTTOM you would get 0 if your Opposite Number chose LEFT and 1.5 if she/he chose RIGHT, and 0 is less than 1.5.]

Q.3 What is the worst possible outcome for you in terms of Points if you chose TOP? [*Answer.* 1. If you chose TOP you would get 1.5 if your Opposite Number chose LEFT and 1 if she/he chose RIGHT, and 1 is less than 1.5.]

Q.4 What is the best possible outcome for your Opposite Number in terms of Points if she/he chose RIGHT? [*Answer.* 2. If he/she choses RIGHT he would get 2 if you chose TOP and 0.5 if you chose BOTTOM, and 2 is more than 0.5.]

Q.5 Suppose you think your Opposite Number is equally likely to choose LEFT or RIGHT. Do you expect to get a higher payoff from choosing BOTTOM or TOP? [(Answer 5) TOP. If you choose TOP you think you will get 1.5 and 1 with equal likelihood, so your expectation is 1.25; if you choose BOTTOM it is only 0.75.]

Ask for help by putting up your hand if you are puzzled. If you are happy, click on OK.

Next, we explain what LIKELIHOOD DECISIONS (L DECISIONS) are, and give some practice on them. In Stages 1 and 2 you will either make L Decisions yourself, or you will interact with participants who do. In the latter case, your ability to make good I Decisions will depend on understanding how L Decisions are made.

L Decisions

In an L Decision (L for Likelihood), you are asked to report your opinion on how likely a certain UNKNOWN FACT is to be true. You do this as a percentage figure, called your REPORT. For example, if your Report is 50 percent, this means you think the chance that the Unknown Fact is true is 1 in 2. Similarly, a Report of 100 means you are certain it is true, and a Report of 0 means you are certain it is not true. We will now give you a practice example in which the Unknown Fact is a simple fact about the

world. (In the main part of the experiment, the Unknown Fact will be which Option your Opposite Number chooses.) Here is the example.

PRACTICE L DECISION 1: How likely do you think it is, on a scale of 0 to 100, that A CARD DRAWN FROM A COMPLETE WELL- SHUFFLED PACK WILL BE A SPADE? Use the mouse to move the pointer on the dial to show your answer. [Answers].

In Stages 1 and 2 each participant who makes L Decisions will be paid for one of them. This payment is called her/his L-PAYMENT. It may be anything up to £3. The L-Payment is determined as follows. You begin with £3. After you have made your Report, an amount gets deducted from your £3 (your L-DEDUCTION) and you get paid what is left. The L- Deduction is calculated by a formula. If the Unknown Fact turns out to be TRUE, then the higher your Report was, the smaller is the Deduction, and the more you are paid. If it turns out to be FALSE, then the lower your Report was, the smaller is the Deduction and the more you get paid. It is important to realize that YOUR EXPECTED L-PAYMENT IS MAXIMIZED IF YOU REPORT YOUR LIKELIHOODS CAREFULLY AND TRUTHFULLY. This payment scheme is known as the Quadratic Scoring Scheme. The formula, and a full explanation of why it benefits you to report your likelihood carefully and truthfully, are available on request after the experiment (ask for the Handout).

Here's one more example. This time, we'll ask you to make a decision for 'fictitious money'. Imagine you have been given £3. We'll ask you to make an L Decision, then tell you whether the Unknown Fact in this L Decision is true or not, and the payment you would finish up with on this decision.

PRACTICE L DECISION 2: On a scale of 0 to 100, how likely do you think it is that THE RIVER AMAZON IS MORE THAN 3000 MILES LONG?

Remember, if you think that it's a toss up whether the Amazon is longer than 3000 miles, then you maximize what you expect to receive by a Report of 50%, if you think the likelihood is 75%, you maximize it by a Report of 75%, and so on. [Answers]. The length of the Amazon is 3900 miles. Your Report was [report]. Your deduction is therefore £[deduction] and your L- Payment is £[payment].

G Decisions

In a G Decision (G for Guess), you are asked to make a Guess about what your Opposite Number's Report in an L Decision was. Let us try out a G Decision. In this one, which is purely for practice, we ask you to imagine that you have an Opposite Number. Imagine that your Opposite Number has just made the following L Decision: 'On a scale of 0 to 100, how likely do you think it is that THE RIVER AMAZON IS MORE THAN 3000 MILES LONG?'. Now here is your G Decision:

PRACTICE G DECISION: Please make your GUESS what your Opposite Number's Report was. [Answers].

In Stages 1 and 2 each participant who makes G Decisions will be paid for one of them. This payment is called her/his G-PAYMENT. The G- Payment rewards participants for the accuracy of their Guesses. This is the scheme on which G-Payments are determined. You begin with an initial sum of money, and for every percentage point your guess is 'out' a deduction is made. But you cannot lose more than your initial sum of money.

The I Decisions of Stages 1 and 2

We come now to the Interactive or I Decisions that you will make in Stages 1 and 2. They represent a type of I Decision that is very common in real life. The BREB group is active in research into discovering how people make I Decisions of this kind. In these I Decisions there are two people, a MOVER and a RESPONDER. The Mover can choose either MOVE or PASS. If she/he chooses Move, the Responder can choose between RESPONSE A and RESPONSE B, and which one he/she chooses affects how well or badly off both finish up. If the Mover chooses PASS, the Responder cannot affect the position of either. Here's an example. Theo has a promising research idea, but lacks the resources to develop it alone. Theo has to decide whether to tell a potential collaborator, Alex, about the idea. If Theo does so, then if Alex decides to collaborate, Theo will benefit. However, once Alex is told the idea, there are ways in which Alex can do better for him/herself by not collaborating. Moreover, in this case, Theo will end up worse off than she/he was originally. In this example, Move is sending the research idea, Pass is not sending it, Response A is collaborating if you are told the idea, and Response

B is not collaborating if you are told it. In the real world we find a great variety of Interactive Decisions which have the form of Mover-Responder problems.

In Stages 1 and 2 you will be asked to make choices in two different Mover-Responder problems, one in Stage 1 and one in Stage 2. You will take your I Decisions for POINTS. At the end of the experiment you will be paid for an I Decision at the rate of £1 per Point. We will explain the payment procedure fully in a minute. Some of you will be Movers and some Responders. Both Movers and Responders may win Points in their I Decisions, but Movers may also in some circumstances lose Points. Which role you play will be determined by your own choices in a lottery. We will now proceed to this lottery.

THE LOTTERY. You will see a display of 10 'nonsense syllables'. The computer has assigned each of them, randomly, a code number between 1 and 80. No two code numbers are the same. You will be asked to choose one syllable. When everyone has made a selection, the participants with the four lowest code numbers will be assigned the Mover's role and the others the Responder's role. You will be shown immediately to which role you have been assigned. Please keep your role to yourself both during and after the experiment. In each of your four I Decisions your Opposite Number will be a different person. You will not learn at any stage who your four Opposite Numbers were. Please choose one 'nonsense syllable'. Please wait for your Role Assignment. [Role is assigned].

II. PLAY STAGE [R subjects]

You have been assigned the Mover role.

Stage 1: General description

This Stage has two ROUNDS. In each Round you face the same Mover-Responder problem, Problem 1. In each Round you will make an I Decision with a different Opposite Number. Your Point Table for Problem 1 is shown above.

Your I Decision is to choose between Move and Pass, and your Opposite Number's is to choose between Response A and Response B. But before you make your I Decision, you will be asked to make an L Decision. The L Decision is to decide on a Report, on the usual dial, of how likely you think it is that your Opposite Number will choose Response A.

In thinking about this, you may like to consider what the procedure will be for people in the Responder role. Each Responder, before making his/her I Decision between Response A and Response B, will be given some information about the L Decisions of participants in the Mover role. However, this information will not include any about YOUR L Decision. Instead, your Opposite Number will be told the average Report of the three Movers other than you. For example, if you report the likelihood R1 and the other Movers report the numbers R2, R3 and R4, your Opposite Number will be told the average of R2, R3 and R4, that is, (R2 + R3 + R4)/3. (Your Report, R1, will be an ingredient in the average figures given to the Responders who are NOT your Opposite Number.)

To summarize, the procedure for you in each Round is as follows.

1. You make your L Decision, a Report about the how likely you think it is that your Opposite Number will choose Response A.

2. Your Opposite Number is told the average of the L Decisions of all Movers except you.

3. You make your I Decision between Move and Pass. At the same time as this your Opposite Number is asked to make his/her I Decision between Response A and Response B.

Nobody will be told anything about anyone else's I Decisions until the end of the experiment. So you will learn nothing about outcomes of your I Decisions or of your L Decisions until the end.

Stage 2: General description

Like Stage 1, Stage 2 consists of two Rounds, in each of which you make an L Decision and an I Decision. In each Round you will have an Opposite Number with whom you have not previously interacted in the experiment. The only difference from Stage 1 is in the Mover-Responder problem, which is a different variant, Problem 2, with a different Point Table.

At the end of the experiment your total payment will be determined as follows. The computer programme will randomly choose one of the four rounds of Stages 1 and 2 as

your L-PAYMENT ROUND, and a different round as your I-PAYMENT ROUND. You will be reminded of your L Decision in your L-Payment Round and shown the actual choice of your Opposite Number in that round and your resulting L-Payment. You will be reminded of your I Decision in your I-Payment Round and shown the choice of your Opposite Number in that round and the number of Points you scored, which may be positive or negative. Your final payment will be the sum of

- your Starting Credit of £4,
- your L-Payment in your L-Payment Round,
- the Points you scored in your I Decision in your I-Payment Round, which may be positive or negative, converted into money at £1 per Point.

* * * * *

When everyone is ready, we will begin Round 1 of Stage 1. You may have to wait one or two minutes for others to be ready. We ask you to be patient. Be sure you have understood the whole procedure, referring to Help if you wish to, before clicking on Continue. Put up your hand if you need any further help. Please look at the Point Table for your I Decision presented above [Problem 1 redisplayed]. You will be asked to say, on a scale of 0 to 100, how likely you think it is that your Opposite Number will choose Response A.

[The programme loops to here. When all have completed rounds 1 and 3 subject reads:] In this Round's I-Decision you have a new Opposite Number. Your Point Table is unchanged, as shown above. Please look at it. In a moment you will be asked to say, on a scale of 0 to 100, how likely do you think it is that your Opposite Number will choose Response A.

[When all have completed round 2 screen announces Stage 2 and reads:] This Stage is just like the last one except that the I Decision is for a different Mover-Responder problem, Problem 2. Your Point Table for Problem 2 is the one shown above. Please look at it carefully. If you wish to compare it with the one for Problem 1, you can view the latter by clicking the button marked Problem 1. In a moment, you will be asked how likely you think it is that your Opposite Number will choose Response A.

YOUR L DECISION FOR THIS ROUND. On a scale of 0 to 100, how likely do you

think it is that your Opposite Number will choose Response A?

If this Round turns out to be your L-Payment Round, you will be paid for your Report in the way we have explained. It therefore pays you to report accurately. Try to make a decision within a few minutes. But do not rush, and ask help if you need it. When everyone has decided, we will proceed to your I Decision. If you need to consult the Table press the 'TABLE' Button [Subject answers]. Please wait for other participants.

[When all E subjects as have received reports:] YOUR I DECISION FOR THIS ROUND. Your Opposite Number has now been told the average Report of all the Movers but you, and is now making her/his I Decision for this round. We remind you that the Point Table for the I Decision is the one shown above. Please look at it carefully. Choose between Move and Pass. Do not rush, but try to make a decision within a few minutes. When everyone has decided, we will proceed to the next Round.

Please wait for other participants. [After all participants have finished this round, the process loops until the end of round 4.]

II. PLAY STAGE [E subjects]

You have been assigned the Responder role.

Stage 1: General description

This Stage has two ROUNDS. In each Round you face the same Mover-Responder problem, Problem 1. In each Round you will make an I Decision with a different Opposite Number. Your Point Table for Problem 1 is shown above.

Your Opposite Number's I Decision is to choose between Move and Pass, and yours is to choose between Response A and Response B. Before your Opposite Number makes her/his I Decision, she has been asked to make an L Decision. This is to make a Report on how likely she thinks it is that you will choose Response A.

In each Round, before you make your I Decision between Responses A and B, you will be asked to make a G Decision. This G Decision is to make a GUESS about your Opposite Number's L Decision – that is, to guess her Report on how likely she thinks it is that you will choose Response A. For example, you might think that she has said you

are pretty likely to choose Response A – you might think she gave this a likelihood of 80%; or you might think she has said you are pretty unlikely to – she gave it a likelihood of 20%. In the former case you would stand to gain most money by answering 80% in your G Decision; in the latter case by answering 20%.

To help you Guess we are going to tell you what actual people in the Mover role HAVE reported to us. We will not tell you what your own Opposite Number reported, but we will tell you the AVERAGE Report of the other three Movers in this Round. That is, if your Opposite Number reported the likelihood R1 and the other Movers report the numbers R2, R3 and R4, you will be informed of the average of R2, R3 and R4, that is, (R2 + R3 + R4)/3.

To summarize, the procedure for you in each Round is as follows.

1. Each of the Movers makes an L Decision, giving a Report of how likely she/he thinks it is that you will choose Response A.

2. You are told the average of the L Decisions of all the Movers except your Opposite Number.

3. You make your G Decision, your Guess about the L Decision of your Opposite Number.

4. You make your I Decision between Response A and Response B. At the same time as this, your Opposite Number is making her/his I Decision between Move and Pass.

Until the end of the experiment, nobody will be told anything about anyone else's I Decisions, and you will be told nothing about your own Opposite Numbers' L Decisions. Hence you will learn nothing about outcomes of your I Decisions or of your G and C Decisions until the end.

Like Stage 1, Stage 2 consists of two Rounds, in each of which you make a G Decision and an I Decision. In each Round you will have an Opposite Number with whom you have not previously interacted in the experiment. The only difference from Stage 1 is in the Mover-Responder problem, which is a different variant, Problem 2, with a different Point Table.

At the end of the experiment your total payment will be determined as follows. The computer programme will randomly choose one of the four rounds of Stages 1 and 2 as your G-PAYMENT ROUND, and a different round as your I-PAYMENT ROUND. You will be reminded of your G Decision in your G-Payment Round. You will then be shown the actual L Decision of your Opposite Number in that round and your resulting G- Payment. You will be reminded of your I Decision in your I-Payment Round and shown the choice of your Opposite Number in that round and the number of Points you scored. Your final payment will be the sum of

- your Starting Credit of £4,
- your G-Payment in your G-Payment Round,
- the Points you scored in your I Decision in your I-Payment Round, which may be positive or negative, converted into money at £1 per Point.

* * * * *

When everyone is ready, we will begin Round 1 of Stage 1. You may have to wait one or two minutes at this point. We ask you to be patient. Be sure you have understood the whole procedure, referring to Help if you wish to, before clicking on Continue. Put up your hand if you need any further help.

Please look at the Point Table above for this Round's I Decision [Problem 1 redisplayed]. You will be asked to make a Guess about what your Opposite Number answered to the question: 'On a scale of 0 to 100, how likely do you think it is that your Opposite Number will choose Response A?'

[When all have completed rounds 1 and 3:] In this Round's I Decision you have a new Opposite Number. Your Point Table is unchanged, as shown above. Please look at it. In a moment you will be asked to make a Guess about what your Opposite Number answered to the question: 'On a scale of 0 to 100, how likely do you think it is that your Opposite Number will choose Response A?'

[When all have completed round 2 screen announces Stage 2 and reads:] This Stage is just like the last one except that the I Decision is for a different Mover-Responder problem, Problem 2. Your Point Table for Problem 2 is the one shown above. Please look at it carefully. If you wish to compare it with the one for Problem 1, you can view

the latter by clicking the button marked Problem 1. In a moment, you will be asked to make a Guess about what your Opposite Number answered to the question: 'On a scale of 0 to 100, how likely do you think it is that your Opposite Number will choose Response A?' Please wait for other participants. The average report of all Movers other than your current Opposite Number to the question 'On a scale of 0 to 100, how likely do you think it is that your Opposite Number will choose Response A?' is [value].

YOUR G DECISION FOR THIS ROUND. Your Opposite Number has just made an L Decision. It was: 'On a scale of 0 to 100, how likely do you think it is that your Opposite Number will choose Response A?' On a scale from 0 to 100, please make your Guess what her/his Report was.

If this Round turns out to be your G-Payment Round you will be paid for this Guess, and the amount will depend on how accurate it is. Try to make a decision within a few minutes. But do not rush, and ask help if you need it. When everyone has decided, we will proceed to your I Decision. If you need to consult the Table press the 'TABLE' Button [Answers].

YOUR I DECISION FOR THIS ROUND. We remind you that the Point Table for the I Decision is the one shown above. Your Opposite Number is now making her/his I Decision for this round. Please choose between Response A and Response B. Do not rush, but try to make a decision within a few minutes. We will then proceed to the next Round. [Subject chooses.]

Please wait for other participants. [After all participants have finished this round, the process loops until the end of round 4.]

PAYMENT STAGE [R Subjects]

You have completed the session. You will learn your payoff when everyone has completed it. Please wait.

The Computer chose Round [number] as your L-Payment Round and Round [number] as your I-Payment Round. The result is as follows:

In Round [number] your Report of the likelihood that your Opposite Number would choose Response A was [amount] and he/she chose RESPONSE (A or B). Your L-

Payment is therefore [amount].

In Round [number] you chose (Move or Pass) and your Opposite Number chose RESPONSE (A or B). The Problem was Problem (1 or 2). Your I-Payment is therefore £[amount].

Your Starting Credit was £4. Thus your total payment will be £[amount].

Thank you for participating in the experiment. Please wait for an experimenter to come to you.

PAYMENT STAGE [E Subjects]

You have completed the session. You will learn your payoff when everyone has completed it. Please wait.

The Computer chose Round [number] as your G-Payment Round and Round [round] as your I-Payment Round. The result is as follows:

In Round [number] your Guess about your Opposite Number's Report of the likelihood you would choose Response A was [amount] and her/his Report was [amount]. Your G-Payment is therefore £[amount].

In Round [number] you chose RESPONSE [A or B] and your Opposite Number chose (Move or Pass). The Problem was Problem [1 or 2]. Your I-Payment is therefore £[amount].

Your Starting Credit was £4. Thus your total payment will be £[amount].

Thank you for participating in the experiment. Please wait for an experimenter to come to you.

REFERENCES

BACHARACH, M. O. L. AND D. GAMBETTA (2001a): "Trust in Signs," in *Trust in Society*, ed. by K. Cook, New York: Russell Sage Foundation.

BACHARACH, M. O. L. AND D. GAMBETTA (2001b): "Trust as Type Detection," in *Deception, Fraud and Trust in Agent Societies*, ed. by C. Castelfranchi, Y.-H. Tan *et al.*, Dordrecht: Kluwer.

BERG, J., J. DICKHAUT AND K. MCCABE (1995): "Trust, Reciprocity and Social History," *Games and Economic Behavior* 10, 122-142.

BOLLE, F. (1995): "Rewarding Trust: An Experimental Study," *Theory and Decision* 25, 83-98.

BOLTON, G. AND A. OCKENFELS (2000): "ERC – A Theory of Equity, Reciprocity and Competition," *American Economic Review* 90, 166-193.

DUFWENBERG, M. (in press): "Marital Investments, Time Consistency, and Emotions", *Journal of Economic Behavior and Organization.*

BRANDTS, J. AND G. CHARNESS (2000) "Hot vs Cold: Sequential Responses and Preference Stability in Experimental Games," *Experimental Economics* 2, 227-238.

CASON, T. AND V. MUI (1998) "Social Influence in the Sequential Dictator Game," *Journal of Mathematical Psychology* 42, 248-265.

COLEMAN, J. (1990): *Foundations of Social Theory*, Harvard: Belkamp.

CROSON, R. T. A. (2000): "Thinking Like a Game Theorist: Factors Affecting the Frequency of Equilibrium Play," *Journal of Economic Behavior and Organization* 41, 299-314.

DAVIS, D. AND C. HOLT (1993): *Experimental Economics*, New Jersey: Princeton University Press.

DUFWENBERG, M. AND U. GNEEZY (2000): "Measuring Beliefs in an Experimental Lost Wallet Game," *Games and Economic Behavior* 30, 163-182.

ENGLE-WARNICK, J. AND R. SLONIM (2001): "Inferring Repeated Game Strategies from Actions: Evidence from Trust Game Experiments," Nuffield College, Oxford, Working Paper 2001-W13.

FALK, A. AND U. FISCHBACHER (1999): "A Theory of Reciprocity," Working Paper 6, Institute for Empirical Research in Economics, University of Zurich.

FEHR, E. AND S. GÄCHTER (1997): "How Effective Are Trust and Reciprocity-based Incentives?" in *Economics, Value and Organisation*, ed. by A. Ben-Ner and L. Putterman, Cambridge: Cambridge University Press.

FEHR, E. AND SCHMIDT, K. M. (1999): 'A Theory of Fairness, Competition and Cooperation', *Quarterly Journal of Economics* 114, 817-868.

FREY, B. AND F. OBERHOLZER-GEE (1997): "The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-out', *American Economic Review* 87, 746-755.

GAMBETTA, D. (1988): *Trust: Making and Breaking Cooperative Relations*, ed. by D. Gambetta, Oxford: Blackwell.

GEANAKOPLOS, J., D. PEARCE AND E. STACCHETTI (1989): "Psychological Games and Sequential Rationality," *Games and Economic Behavior* 1, 60-79.

GLAESER, E., D. LAIBSON, J. SCHEINKMAN AND C. SOUTTER (2000): "Measuring Trust," *Quarterly Journal of Economics* 115, 811-846.

GUILKEY, D. K. AND J. L. MURPHY (1993): "Estimation and Testing in the Random Effects Probit Model," *Journal of Econometrics* 59, 301-317.

GÜTH, W., S. HUCK AND W. MÜLLER (in press): "The Relevance of Equal Splits: On a Behavioral Discontinuity in Ultimatum Games," *Games and Economic Behavior.*

HARDIN, R. (1991): 'Trusting Persons, Trusting Institutions' in *Strategy and Choice*, ed. by R. Zeckhauser, Cambridge: MIT Press.

HARGREAVES HEAP, S., M. HOLLIS, B. LYONS, R. SUGDEN AND A. WEALE (1992): *The Theory of Choice*, Oxford: Blackwell.

HAUSMAN, D. (1998): "Fairness and Trust in Game Theory," London School of Economics, mimeo.

HAUSMAN, D. (1998a): 'Trust, Preference, and Interest," London School of Economics and University of Wisconsin, Madison, mimeo.

HIRSCHMAN, A.O. (1988): "Against Parsimony. Three Easy Ways of Complicating Some Categories of Economic Discourse," *American Economic Review Papers and Proceedings* 74,

88-96.

Hollis, M. (1998): *Trust within Reason* Cambridge: Cambridge University Press.

Hume, D. (1740/1978): *Treatise on Human Nature*, Oxford: Clarendon Press.

Jussim, L. (1986): "Self-fulfilling Prophecies: A Theoretical and Integrative Review," *Psychological Review* 93, 429-445.

Kelly, H. and A. Stahelski (1970): "Social Interaction Basis of Cooperators' and Competitors' Beliefs about Others," *Journal of Personality and Social Psychology.* 16, 66-91.

McKelvey, R. and T. Palfrey (1992): "An Experimental Study of the Centipede Game," *Econometrica* 60, 803-836.

Orbell, J. M. and Dawes, R. M. (1991): "A 'Cognitive Miser' Theory of Cooperator's Advantage," *American Political Science Review* 85, 515-528.

Pelligra, V. (2000): 'Goldfish and Game Theory: A Problem of Trust', School of Economic and Social Studies, University of East Anglia, mimeo.

Pettit, P. (1995): "The Cunning of Trust," *Philosophy and Public Affairs* 29, 202-225.

Putnam, T. D. with R. Leonardi and R. Y. Nanetti (1993): *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton: Princeton University Press.

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

Rigdon, M., K. McCabe and V. Smith (2000) 'Positive Reciprocity and Intentions in Trust Games', University of Arizona, mimeo.

Schotter, A., K. Weigelt and C. Wilson (1994): "A Laboratory Investigation of Multiperson Rationality and Presentation Effects," *Games and Economic Behavior* 6, 445-468.

Schotter, A., A. Weiss and I. Zapater (1996): "Fairness and Survival in Ultimatum and Dictator Games," *Journal of Economic Behavior and Organization* 31, 37-56.

Selten, R. (1967): "Die Strategiemethode zur Erforschung des Eingeschränckt Rationalen Verhaltens in Rahmen eines Oligopolexperiments', in *Beitraägge zur Experimentellen Wirtschaftforschung*, ed. by H. Sauermann, Tübingen: J. C. B. Mohr.

Shafir, E. and Tversky, A. (1992): "Thinking through Uncertainty: Nonconsequential Reasoning and Choice," *Cognitive Psychology* 24, 449-474.

STAHL, D. AND P. WILSON (1995): "Experimental Evidence on Players' Models of Other Players", *Journal of Economic Behavior and Organization* 25, 309-327.

WRIGHTSMAN, L. S. (1966): "Personality and Attitudinal Correlates of Trusting and Trustworthy Behaviors in a Two-person Game," *Journal of Personality and Social Psychology* 4, 328-332.

ZIZZO, D. J. (2000): *Relativity-Sensitive Behaviour in Economics*, University of Oxford, doctoral thesis.

FOOTNOTES

1. The latter term is Hirschmann's (1988). Others are the 'trust mechanism' (Hausman 1998) and 'positive responsiveness' (Bacharach and Gambetta 2001).

2. The same is not true of C in a standard PD. Here the only strategies for each player are C and D. If we identify C with T and F, and D with W and V then, considering without loss of generality the row player, (1) and (3) hold, but (2) fails because she is made worse, not better, off by T if the column player plays F.

3. In the sequential version, the only subgame perfect equilibrium is (W, V); in the normal form V weakly dominates F, and the only trembling-hand perfect equilibrium is (W, V). In '$2 \times \infty$' versions in which E chooses y in $[0, 1]$, if R chooses T her payoff gain over W is positive for $y = 1$, negative for $y = 0$, and increasing in y, but since E's payoff is decreasing in y it is dominant for E to choose $y = 0$, and hence for R to choose W. In the '$\infty \times \infty$' version iterated dominance similarly gives zero degrees of fulfilment and of trusting similarly.

4. The hypothesis that experimental subjects assimilate the decision problem faced in the laboratory to related but different real-life ones.

5. Assuming a modest degree of iterated knowledge of rationality and the game.

6. Hume (1740/1978) maintains that people do not just wish to be well thought of, but also to have particular qualities that others admire, such as trustworthiness. Hausman notes Oliver Twist's reaction to the thought that Mr Brownlow, who has trusted him, will be told that Oliver has stolen his books. His distress illustrates the deep importance to us of being thought trustworthy by those we respect, and of not losing this good opinion.

7. Bolton and Ockenfels' (2000) model also has this property.

8. Roughly, K is R's perception of E's benefit from her action beyond what equitability calls for, normalized by R's scope for affecting E's payoff.

9. A determinate prediction involves the absolute size of the material payoffs x and z. Rabin leaves it entirely open how material payoffs might be calibrated against the utility from reciprocating kindness.

10. Unlike the Falk-Fischbacher model, Rabin's model does not imply that trust responsiveness

is negative, but makes no prediction about it, since the model is silent on the relation, if any, between t^* and f^{**}. If t^* increases with f^{**}, then a rise in f^{**} could raise f by pushing up t^* enough to satisfy (6). We might in fact expect t^* to increase with f^{**}: a rise in f^* could well raise t by raising R's expected payoff from T, and then t^* increases with f^{**} provided that E has a model of R which recognizes this. However, this mechanism for raising t^* could not lead E to fulfil within the spirit of Rabin's theory, since in it E believes that R is motivated to play T purely for personal gain, not out of kindness. Although Rabin's theory thus effectively rules out trust responsiveness, it does imply that a sufficient level of f^{**} is necessary condition for fulfilling. This is because it is an equilibrium theory, and in equilibrium $f^{**} = f^* = f$, so fulfilling implies positive f^{**}.

11. Rigdon *et al.* suggest that the greater the opportunity cost to R of trusting, the more will E be inclined to fulfil; the thought is that a trusting act is kinder the more you have to give up to do it. E's perception of R's cost of trusting is $f^{**}(a-w)+(1-f^{**})(a-y)$. For any f^{**} this decreases as $-a$ rises, and the Rigdon effect of a rise in 'need' $-a$ is therefore a fall in f. The part of the effect due to the second cost term, relating to R's reduced exposure, dwindles as f^{**} grows.

12. Conversely, trust responsiveness may operate negatively when sympathy or respect are lacking, because the trustee may then interpret a high f^* as 'taking him for granted'.

13. Dufwenberg and Gneezy's (2000) design included (ii) and (iii).

14. It has been argued (Croson 2000) that eliciting beliefs may change behavior. We think this is a serious possibility, and that the existence and size of such possible framing effects should be carefully studied, but that the only route to empirical knowledge about questions such as the present one is by elicitation, with due attention to incentive compatibility. Also, some of the most important applications of the notion of trust responsiveness, including some policy applications we mention later, are to situations in which, as in our design, trustees' attention is drawn to the truster's confidence; for these appplications, elicitation may increase 'external validity'.

15. An alternative design would use a summary statistic of R subjects in other sessions, but we judged that the statements of co-sessioners would be perceived as more 'relevant'.

16. Labelling the players R1, ... , R4, E1, ... , E4, R1 played in turn with E1, E2, E3, E4; R2 played in turn with E2, E3, E4, E1; R3 with E3, E4, E1, E2; and R4 with E4, E1, E2, E3.

17. Subjects' statements and guesses were made as integers between 0 and 100. A subject stating s received £3$[1-(1-0.01s)^2]$ if her coplayer chose F and £3$(1-0.01s^2)$ if he chose V. A subject guessing g received £3 if g was correct, and 30 pence less for each unit of error, subject to nonnegativity of the payment (so she received nothing if her guess was 10 or more percentage points out).

18. Subjects R6, R8, R11 and R25 displayed such behavior. One possibility is that they were trying to hedge their game payoff risks.

19. The simplest error model is as follows. With probability $1-e$, R chooses according to the theory, and with probability e at random. Then $t = u + .5e$, where u is the fraction of tasks in which $s \geq f_{\mathsf{crit}}$. We have, in round figures: $t = 0.5, u = 0.3$ (since of the fraction 0.5 of T choices, about 0.6 were correct according to the theory); hence e is about 0.4. Error rates found in other studies are up to about 0.25, so to reconcile the behavior of R players with the theory one needs to introduce something that lowers f_{crit}. One possibility is risk-preference; another is utility from trusting.

20. If y is E's transfer, then any E player for whom all-in utility is positive at $y = 0$ and negative at $y = 1$ will transfer something in the fractional fulfilment game but choose V in the BTG.

21. The measure ctr is rough and ready in two ways. It is the gradient of f on g in the regression of g on f rather than f on g, and the latter regression estimates the 'linear probability model' for Pr(F), which at best approximates the *a priori* requirements for such a model.

22. The classes are mostly of fair size (17, 23, 28, 27, 11, 18, 10, 13, 3, 10), but aggregating over the three the variants may have induced bias in the gradient estimate. For example, a positive gradient in the aggregate relationship is consistent with zero trust responsiveness but a tendency for kindness and/or need to produce both high g and high f.

23. We also experimented with other variables not included in the Table 6 equation, including a round counter and a dummy for treatment order, none of which showed any significance.

24. One subject had to leave during a session owing to a computer failure, possibly contaminating the data of others who observed him leave, and two subjects in another session turned out to be a couple.

25. The estimated equation implies that adding a year to the age of a 25 year old male graduate

who plays KTG and guesses 0.25 raises his F probability by 3.1 percentage points.

26. What we and Rabin call kindness might, for instance, be perceived in our BTGs not as kindness but as utilitarianism, since the utilitarian objective is also maximized by T when $f^* = 1$.

27. A conceivable counter-motive is hedging. However, an E player who cares only about material payoff and chooses V has no risk to hedge. The same goes for an E player who plays F out of principle. A trust-responsive E with high f^{**} might perhaps choose F expecting secondary payoff from (T, F), but report a lower f^{**} to increase his accuracy payment in case he is wrong about s. Such hedging (and likewise the reverse case of trust responsiveness, low f^{**}, V and high g) biases the data towards falsely rejecting trust responsiveness.

28. The distribution of guesses has concentration lines at the focal levels 0.25, 0.50 and 0.75: 31 choices out of 160 had one these values, while a normal fitted to the observed distribution predicts 5.1 out of 160.

29. The round-robin design means that when E plays his second, third and fourth games he can make inferences from the new report about the statement of his current coplayer. For example, if the old mean was 67 and the new mean is 57 he might infer that his current coplayer returned a statement of between 30 and 100. In theory, realizing this might give an R player a reason to misrepresent; however, there is no obviously advantageous way to do so, and misrepresenting is strongly opposed by the incentive for accuracy in reporting confidence. In any case, E has no assurance that his current coplayer always makes the same statement. The inferences that could be drawn by E are pretty diffuse. They would be even more diffuse if the groups of R subjects had been larger than four, but we preferred to keep the number low to get variance in the report and so in E players' beliefs.

30. Although the correlation is highly significant, a single experiment with 80 or so subjects does not provide enough power to show trust responsiveness by instrumenting g on r. If the correlation is corroborated in further studies, however, this will provide a second, quite different way of showing that causation runs from f^{**} to f.

31. Advertisment by the Australian Bureau of Statistics, July 2001.

32. The point of advertisements such as this one by British Gas may be that if the truster (the customer) believes the message she will think the trustee (BG) is a trust-responsive type

with a high f^{**} and so she will expect fulfilment.

33. Dufwenberg (in press) has argued that forward induction reasoning of this kind may explain why a trust-responsive spouse with a financial incentive to divorce may stay in a marriage.