Theory and Decision (2007) 63:349–388
DOI 10.1007/s11238-007-9043-5

MICHAEL BACHARACH, GERARDO GUERRA
and DANIEL JOHN ZIZZO

THE SELF-FULFILLING PROPERTY OF TRUST: AN EXPERIMENTAL STUDY

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 discriminate 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 they do so only in conjunction with trust responsiveness.

KEY WORDS: trust, trust responsiveness, kindness, need to trust, belief elicitation

JEL CLASSIFICATION CODES: C72, C92, D84

1. INTRODUCTION

Today few doubt the importance of trust and trustworthiness as explanatory factors in economic behaviour. 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*, the tendency to fulfil trust because you believe it has been placed in you. Similar concepts are the 'trust mechanism' (Hausman, 1998) and 'positive responsiveness' (Bacharach and Gambetta, 2001a,b). 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 behaviour goes directly against material incentives.

In Section 2 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 3 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 'behavioural' 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 4 reports results, Section 5 discusses the formation of subjects' beliefs and the direction of causality between rates of fulfilment and other variables, and Section 6 summarizes and draws some implications for policy and theory.

2. TRUST AND TRUST RESPONSIVENESS

2.1. *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 I, and whose coefficients satisfy the three inequalities:

$$y < a \quad \text{(Exposure)}, \tag{1}$$

$$a < w \quad \text{(Improvement)}, \tag{2}$$

$$x < z \quad \text{(Temptation)}. \tag{3}$$

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) and

TABLE I
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

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 (BTG) 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 as specified in Table I (where $X = \mathrm{F}$) and chooses T because she expects E to choose X. 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'.[1]

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

$$b < x. \quad \text{(Mutual Gain)} \tag{4}$$

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 Prisoner's 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.

2.2. *Behaviour in trust games*

In any BTG, the standard prediction based on the assumption of purely self-interested agents is (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.[2] Yet in all forms of trust game which have been studied in the laboratory, the standard prediction is quite systematically violated, often to a significant degree (although in some experiments less than in others).

The experimental literature on the BTG and its variants includes Berg et al. (1995); Bolle (1995); Fehr and Gächter (1997); Dufwenberg and Gneezy (2000), and Schotter and Sopher (2006) among others. The central result is that half or more of the 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% on average, 90% of them a positive amount, and nearly 50% of *E* subjects returned more than the amount *R* transferred. Because these are not repeated games, trusting and fulfilling behaviour cannot be explained as rational strategies for getting future rewards.

2.3. *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, i.e. that experimental subjects assimilate the decision problem faced in the laboratory to related but different real life ones.

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 $U_i = \Sigma_j u_j$.

Formal theories of trust responsiveness represent the disposition by a special kind of payoff transformation. A 'psychological game' (Geanakopolos 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 I, *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, 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 between strangers the case for expecting equilibrium behaviour is weak. Second, 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 creates an identification problem. Very often the same behaviour 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.

2.4. *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, 1984; 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 the idea of 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). This may be especially so of those

he respects (Hausman, 1998).[3] Hausman adds that the more certain someone is of this good opinion, the more strongly he will wish to keep it. So, 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.

While our focus is on trust games, we recognize that trust responsiveness may be an aspect of a more general psychological mechanism that makes agents sensitive to higher-order beliefs. Blount (1995) provides suggestive evidence of such belief dependence for the case of ultimatum games. Charness and Rabin (2002) propose a theory that is consistent with trust responsiveness but applicable also in contexts, such as dictator games, that are not trust games.

2.5. *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 $v \geqslant u$ and $= -\beta(u-v)$ if $v \geqslant u$, 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

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

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. Another model with this property can be found in Bolton and Ockenfels (2000).

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

$$k(t, f^*) = \frac{v(t, f^*) - \bar{v}(f^*)}{v^h(f^*) - v^l(f^*)}, \tag{5}$$

where $v^h(f^*)$, $v^l(f^*)$ and $\bar{v}(f^*)$ 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. If the Mutual Gain condition (4) holds, R is Rabin-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. A determinate prediction would involve the absolute size of the material payoffs x and z. The theory leaves open how material payoffs might be calibrated against the utility from reciprocating kindness. The mutual gain condition is shown in the Appendix to be

$$z - x < (t^* - 0.5)/t^*. \tag{6}$$

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 (2001) 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

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

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^{**}.[4]

2.6. *Attitudinal theories*

We shall call any theory of trust and fulfilment in which the trustee's choice depends on how favourably 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 neighbour 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.

Charness and Rabin (2002) find that subjects exhibit a special concern for those less well off.

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 McCabe et al. (2003) and Pelligra (2000).[5]

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 2.4, 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. We label the hypothesis that sympathy, respect and other pro attitudes strengthen trust responsiveness the *Interaction Hypothesis*. 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'.

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.

3. DESIGN

3.1. *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; and (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.'[6] We comment on these features in turn.

3.1.1. *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 II. The entries represent gains and losses of money in units of £1.

In the Kind Trust Game (KTG) choosing T has positive Rabin kindness. The Gratuitous Trust Game (GTG), introduced

TABLE II
Three variants of the Basic Trust Game

	Gratuitous (GTG) Trustee (E)		Kind (KTG) Trustee (E)		Needy (NTG) 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 Pelligra (2000), has the same parameters as the KTG except that $b = 3$, making T have zero Rabin kindness when R fully expects F. And for every level of confidence f^*, intuitively kindness is lower in the GTG than in the other two variants. If $f^* = 0.5$, for instance, playing T raises E's expected payoff by 3.75 in KTG and NTG, but only by 0.75 in GTG. This intuition is not captured by Rabin's measure, because it is insensitive to the absolute difference R can make to E's payoff. 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.

3.1.2. *Simultaneous moves*

Subjects chose actions simultaneously, rather than sequentially, in our trust games. An advantage of this is that it provides data on E's preferences at unreached nodes. Another arises from the nature of trust responsiveness. To test it 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, assuming minimal rationality, 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

$$f_{\text{crit}} = \frac{a - y}{w - y}, \tag{8}$$

which is equal to 0.25 in NTG and to 0.5 in GTG and KTG.

3.1.3. *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.[7] 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 E_y. 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). Schotter and Sopher (2006) measure beliefs directly for a fractional version of the game where there can be partial fulfilment, but our version is more manageable and purer in its predictions.

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 E_y, 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.

3.1.4. *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, 2001) 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.[8]

3.2. *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. Each subject played four rounds of the BTG, one with each of the subjects in the other role.[9] 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; there proved to be no significant order effects. In all there were

16 rounds of KTG, and 12 each of GTG and NTG. 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 NTG, and thus 160 values for each behaviour 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, 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 'point 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: each R chose T or W, and simultaneously her coplayer chose F or V.

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^{**}. To summarize: *s measures f^*; g measures f^{**}.*

We will discuss questions about the likely accuracy of these measures in due course. 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' behaviour 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 commonly used quadratic scoring rule (e.g., Huck and Weizsäcker, 2002), and guesses according to a triangular scheme.[10] Sessions averaged about 55 min; 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.

3.3. *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 $\partial f/\partial f^{**}$, the gradient of the fulfilment function with respect to confidence-perception, the *coefficient of trust responsiveness*. Of particular interest are seven 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 2, contradicting the standard prediction 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 E's trust responsiveness 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 2 led us to expect to find support for H1, H2, H3 and H4 and to have open minds about H5, H6, and H7.

4. RESULTS

4.1. *Commentary on the data*

Table III shows some summary statistics for the behaviour of trusters and trustees.

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 $p < 0.01$. Even in the GTG, where R might well think E has a weak motive to play F because E does not perceive T as kind, 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.

There is a striking variability both across and within subjects in the statement s, the elicited value of an R player's expression of confidence. Writing s_i for the statement of the

TABLE III

Means and standard deviations of observed variables

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

ith R player, and $\bar{s}_i$ for its within-subject mean, the standard deviations of $\bar{s}_i$ are 0.26, 0.28 and 0.28 in GTG, 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. The empirical distribution of subjects' beliefs about others' choices in the BTG certainly looks very unlike an equilibrium of a BTG modelled as a psychological game. The standard assumptions of models of this kind (Bacharach and Gambetta, 2001b; Dufwenberg, 2002) imply that in any equilibrium there is a common distribution over choices and a common point estimate f^*.

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 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 $s \geqslant f_{\text{crit}}$ and W only if $s \leqslant f_{\text{crit}}$, 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.[11]

We turn to the behaviour of E players. It comes as no surprise that H1 is strongly falsified. The overall rate of fulfilling 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.[12]

We can now begin to 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. The mean values of the guess g in GTG, KTG and NTG were 0.29, 0.43 and 0.39, respectively. Recall that g is a measure of f^{**}. So 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. The data for trustees also show that a number of subjects' guesses closely tracked the report. The Pearson correlation coefficient of r and g across all tasks is 0.47. In Subsection 5.2 we shall look in more the detail at how guesses were influenced by the incoming report.

Were subjects trust responsive? An initial measure of overall trust responsiveness is *crude trust responsiveness* $\text{ctr} = (\bar{g}_F - \bar{g}_V)/\bar{g}_V$, where $\bar{g}_F$ (resp. $\bar{g}_V$) is the mean value of g in the subset of subjects who chose F (resp. V).[13] The magnitude ctr equals 0.68 over the whole sample, *prima facie* evidence in favour 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

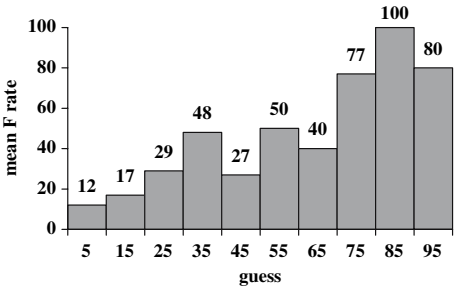

Figure 1. Variation of fulfilling rate with guess.

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.[14] The next subsection presents a more refined analysis, which resolves this and other questions not answered by our preliminary inspection.

4.2. *Econometric analysis*

In the last subsection we drew some preliminary conclusions from the raw data through some statistical measures. These measures decisively rejected H1, and appeared to provide some support for H3, H4 and H6.

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

$$f = \Phi(\beta' x),$$

where $\Phi(y)$ denotes the probability that a standard normal variate is less than y. We began by estimating a general 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 5%, and reintroducing eliminated variables to test for revivals of significance. We did this to minimize leaving out important variables that were initially not significant. Table IV shows

TABLE IV

Fulfilment function: estimate of general model

Variable	Coefficient	S.E.	p
const	−0.842	0.681	0.216
g_G^{a}	0.018	0.009	0.038
g_K^{a}	0.022	0.008	0.008
g_N^{a}	0.027	0.009	0.004
sus^{b}	0.838	0.424	0.048
GTG^{c}	−0.253	0.479	0.597
KTG^{c}	−0.326	0.534	0.542
$male^{d}$	−0.146	0.358	0.684
$dage^{e}$	0.099	0.056	0.079
$grad^{f}$	−0.547	0.426	0.198
hum^{f}	0.170	0.726	0.815
sci^{f}	0.493	0.714	0.490
$socsci^{f}$	0.088	0.678	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%).
[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. Eighty-six percent of subjects were students, of whom 42, 27, 34, and 39% were of these categories.

the result of estimating the general model.[15] 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.[16] The most promising demographic variable is *dage*, the subject's age as a deviation from the mean over subjects, which is significant at 10%.[17] The dummies for the variants are notable for their lack of significance.

The broad pattern of explanation that emerges from the general regression persisted with only minor variations throughout the elimination process. The only 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

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

Figure 2 shows (9) for normal (non-*sus*) 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

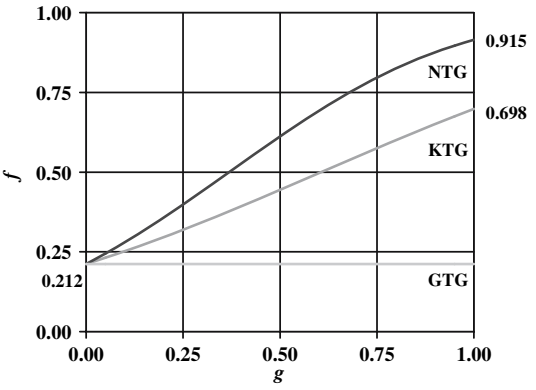

Figure 2. Fulfilment as a function of guess in the three variants.

TABLE V
Fulfilment function: estimate of final model

Variable	Coefficient	S.E.	p	$f'(0.5)$
Const	−0.800	0.250	0.001	–
g_K^{a}	0.013	0.005	0.007	0.522
g_N^{a}	0.022	0.007	0.002	0.831
sus^{b}	0.909	0.343	0.008	0.909

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

demographic variables matter. The estimated rates of trust responsiveness in the two variants with kindness are very substantial.

The probit form means that the coefficients in (9) cannot be identified with the marginal effects with respect to the regressors. Table V displays the coefficients of (9) together with their standard errors and p values, and the estimated coefficients of measured trust responsiveness $\partial f/\partial g$ at $g = 0.5$.

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 0.70 and 0.09; the hypothesis of equal variance yields $F(39, 120) = 7.66$ and $p < 0.01$. The estimation method must take this into account, or else the dataset 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).

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 iterative estimation method resulted in significantly positive slopes for KTG and NTG, with the slope for GTG not significantly different from 0. Furthermore, the test for equality of the coefficients of NTG and KTG, against the alternate that the former exceeds the latter, resulted in a p value of 0.05. We will discuss in the next section how having positive and different slopes helps in determining causality.

The econometric analysis confirms and extends the results of the preliminary 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.

5. 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.

5.1. *Interpretation of the estimated model*

5.1.1. *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 4 lends further support. First, it establishes a necessary condition for trust responsiveness, that f is positively associated with f^{**}. The evidence of positive slopes

in some variants, together with a difference in slopes for different variants, 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 for all f independent of variant). This in turn implies a single conditional distribution of f given g for all g, and so a single curve for all three variants. As the double-projection theory is inconsistent with the separation of the three variant fulfilment functions, the fact that we find separation in our data is evidence against the double-projection hypothesis.

5.1.2. *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'. 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 $f^* = 1$.

5.2. *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, if maybe approximately (these points are discussed more in Bacharach et al. 2001, footnotes 27 and 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.[18]

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,[19] 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 $p < 0.01$. 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 4.2). 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 $p < 0.01$.

6. CONCLUDING REMARKS

6.1. *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 behaviour 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 to 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; and (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.

6.2. *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 more than one perception of that act and hence to more than one attitude to it, with different effects on choice: careful exploration may be needed to disentangle 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."[20] 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".[21] 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 (see also Dufwenberg, 2002).[22] 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 behaviour, and the agent slides down the fulfilment function.

ACKNOWLEDGEMENTS

Michael Bacharach died on August 12, 2002. The experimental instructions and raw data may be found in Bacharach et al. (2001) and at http://www.uea.ac.uk/~ec601/Appendtr2.pdf. We would like to thank Abigail Barr, Andrew Colman, Miguel Costa-Gomes, Jim Engle-Warnick, Diego Gambetta, David Hendry, and an anonymous referee 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.

APPENDIX. 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^*, $v^h(f^*) = T$, $v^l(f^*) = \mathrm{W}$. Since $v(t, f^*) = tv(\mathrm{T}, f^*) + (1-t)v(\mathrm{W}, f^*)$,

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

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

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

$$V = v + K^*(1 + L), \tag{A.1}$$

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 $\mathrm{K}^* = \mathrm{K}(t^*, f^{**})$. Then E chooses F only if his secondary utility is positive, which requires $t^* > 0.5$, and in this case if and only if $\mathrm{V(F)} > \mathrm{V(V)}$, that is, from (A.1),

$$t^*x + (1-t^*)b + 1.5(t^* - 0.5) > t^*z + (1-t^*)b + 0.5(t^* - 0.5), \quad \text{or} \quad t^* - 0.5 > (z-x)t^*.$$

NOTES

1. 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.
2. 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 gives zero degrees of fulfilment and of trusting similarly.

3. 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.
4. 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, the rise in t^* would not lead E to fulfil on 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 implicitly rules out trust responsiveness, it does imply that a sufficient level of f^{**} is a necessary condition for fulfilling. This is because it is an equilibrium theory, and in equilibrium $f^{**}=f^*=f$, so fulfilling implies positive f^{**}.
5. McCabe 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 McCabe 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.
6. Dufwenberg and Gneezy (2000) design included (ii) and (iii).
7. Eliciting beliefs may change behavior in linear public good and Prisoner's Dilemma experiments (Croson, 2000), but Guerra and Zizzo (2004) find no difference in trusting and fulfilling rates between comparable treatments with and without belief elicitation when the BTG is played.
8. 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.'

9. Labeling 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.
10. 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).
11. 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 $s \geqslant f_{\text{crit}}$. 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.
12. 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.
13. The measure ctr is rough 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.
14. The classes are mostly of fair size (17, 23, 28, 27, 11, 18, 10, 13, 3, 10). The across-variants Pearson correlation coefficient between the mean F rate and the guess is 0.41 ($p<0.01$, two-tailed).
15. We also experimented with other variables not included in the Table IV equation, including a round counter and a dummy for treatment order, none of which showed any significance.
16. 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.
17. 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. We use *dage*, rather than just age, because doing otherwise may create spurious significance of the variable: for every single observation it would take a value of 18 or above, and so it might do the work of the intercept in the estimation.
18. A possible concern is that by giving the E player a report we are biasing his stated value g of his unobserved f^{**} towards actual f^{*}. But this would only affect the relationship between f^{**} and t, the

propensity to trust, and our interest is not this relationship but that between f^{**} and f.

19. 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.
20. Advertisement by the Australian Bureau of Statistics, July 2001.
21. 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.
22. Dufwenberg (2002) 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.

REFERENCES

Bacharach, M.O.L. and Gambetta, D. (2001a), Trust in signs, in Cook, K. (ed), *Trust in Society*, Russell Sage Foundation: New York.

Bacharach, M.O.L. and Gambetta, D. (2001b), Trust as type detection, in Castelfranchi, C., Tan, Y.-H., et al. (eds.), *Deception, Fraud and Trust in Agent Societies*, Kluwer: Dordrecht.

Bacharach, M.O.L., Guerra, G. and Zizzo, D.J. (2001), Is trust self-fulfilling? An experimental study, Discussion Paper no. 76, Department of Economics, University of Oxford.

Berg, J., Dickhaut, J. and Mccabe, K. (1995), Trust, reciprocity and social history, *Games and Economic Behavior* 10, 122–142.

Blount, S. (1995), When social outcomes aren't fair: the effect of causal attributions on preferences, *Organizational Behavior and Human Decision Processes* 63, 131–144.

Bolle, F. (1995), Rewarding trust: an experimental study, *Theory and Decision* 25, 83–98.

Bolton, G. and Ockenfels, A. (2000), ERC – A theory of equity, reciprocity and competition, *American Economic Review* 90, 166–193.
Charness, G. and Rabin, M. (2002), Understanding social preferences with simple tests, *Quarterly Journal of Economics* 117, 817–869.
Coleman, J. (1990), *Foundations of Social Theory*, Belkamp: Harvard.
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.
Dufwenberg, M. (2002), Marital investments, time consistency, and emotions, *Journal of Economic Behavior and Organization* 48, 57–69.
Dufwenberg, M. and Gneezy, U. (2000), Measuring beliefs in an experimental lost wallet game, *Games and Economic Behavior* 30, 163–182.
Falk, A. and Fischbacher, U. (2001), Distributional consequences and intentions in a model of reciprocity, *Annales d'Economie et de Statistique* 63–64, 111–129.
Fehr, E. and Gächter, S. (1997), How effective are trust and reciprocity-based incentives? in Ben-Ner, A. and Putterman, L. (eds.), *Economics, Value and Organisation*, Cambridge University Press: Cambridge.
Fehr, E. and Schmidt, K. M. (1999), A theory of fairness, competition and cooperation, *Quarterly Journal of Economics* 114, 817–868.
Frey, B. and Oberholzer-Gee, F. (1997), The cost of price incentives: an empirical analysis of motivation crowding-out, *American Economic Review* 87, 746–755.
Gambetta, D. (ed.) (1988), *Trust: Making and Breaking Cooperative Relations* Blackwell: Oxford.
Geanakopolos, J., Pearce, D. and Stacchetti, E. (1989), Psychological games and sequential rationality, *Games and Economic Behavior* 1, 60–79.
Glaeser, E., Laibson, D., Scheinkman, J. and Soutter, C. (2000), Measuring trust, *Quarterly Journal of Economics* 115, 811–846.
Guerra, G. and Zizzo, D.J. (2004), Trust responsiveness and beliefs, *Journal of Economic Behavior and Organization* 55, 25–30.
Guilkey, D.K. and Murphy, J.L. (1993), Estimation and testing in the random effects probit model, *Journal of Econometrics* 59, 301–317.
Hardin, R.: (1991), Trusting persons, trusting institutions, in Zeckhauser, R. (ed.), *Strategy and Choice*, MIT Press: Cambridge.
Hargreaves Heap, S., Hollis, M., Lyons, B., Sugden, R. and Weale, A. (1992), *The Theory of Choice*, Blackwell: Oxford.
Hausman, D. (1998), Fairness and trust in game theory, Mimeo, London School of Economics.
Hirschman, A.O. (1984), 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 University Press: Cambridge.

Huck, S. and Weiszäcker, G. (2002), Do players correctly estimate what others do?, *Journal of Economic Behavior and Organization* 47, 71–85.
Hume, D. (1740[1978]), *Treatise on Human Nature*, Clarendon Press: Oxford.
Jussim, L. (1986), Self-fulfilling prophecies: a theoretical and integrative review, *Psychological Review* 93, 429–445.
McCabe, K., Rigdon, M. and Smith, V. (2003), Positive reciprocity and intentions in trust games, *Journal of Economic Behavior and Organization* 52, 267–275.
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, Mimeo, School of Economic and Social Studies, University of East Anglia.
Pettit, P. (1995), The cunning of trust, *Philosophy and Public Affairs* 29, 202–225.
Putnam, R.D., Leonardi, R. and Nanetti, R.Y. (1993), *Making Democracy Work: Civic Traditions in Modern Italy*, Princeton University Press: Princeton.
Rabin, M. (1993), Incorporating fairness into game theory and economics, *American Economic Review* 83, 1281–1302.
Schotter, A. and Sopher, B. (2006), Trust and trustworthiness in inter-generational games: an experimental study of inter-generational advice, *Experimental Economics* 9, 123–145.
Schotter, A., Weiss, A. and Zapater, I. (1996), Fairness and survival in ultimatum and dictator games, *Journal of Economic Behavior and Organization* 31, 37–56.
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, Doctoral thesis, University of Oxford.

Address for correspondence: Dr. Daniel John Zizzo, School of Economics, University of East Anglia, Norwich, NR4 7TJ, UK. Phone: +44-1603-593668; Fax: +44-1603-250434; E-mail: d.zizzo@uea.ac.uk
Michael Bacharach, BREB Research Unit, Department of Economics, University of Oxford, Oxford, UK.
Gerardo Guerra, Wolfson College, University of Oxford, Oxford, UK; UDEM Graduate School of Business, University of Monterrey, Monterrey, Mexico.