Plausible deniability and its effects on trust-fulfilling in the basic trust game

James Rossington

Experimental Economics II

I. Introduction

This experiment investigates the effects on trust-fulfilling, of introducing "plausible deniability" into the basic trust game via an *act of chance* which will violate trust on a subject's behalf with some given probability. In section two, the background of moral "wiggle room" and plausible deniability is given in the context of the Dana, Weber and Kuang (2007) experimental paper and the motivation for this experiment is put forward. Section three explains the precise experimental design and procedures which were used and section four sets out the null and alternative hypotheses on which the analysis will focus. Section five presents the findings from the experiment, both raw statistics and the results from a probit regression on both the decision to trust and the decision to fulfil trust. Finally, section six and seven contain a general discussion of the results and limitations of the study and present some suggestions as to the direction in which future research on this topic could be taken.

II. Background

It is a general finding of Economic experiments that subjects display an apparent concern for others' welfare, beyond any scope for reputation or reciprocity. This is most starkly observable in the simple dictator game setting, in which past experimenters have found that a majority of subjects will give some positive amount of their endowment, even when it is not in their best interests to do so (Camerer, 2003, chapter 2). Dana, Weber and Kuang (2007), henceforth DWK, consider the idea that generosity in these and in other settings may simply be a desire to appear fair to others, as opposed to an explicit desire for fair outcomes. Whilst many theories of social preference exist which suggest that giving is dependent on the final outcomes of the game itself[1], DWK suggest that alternative motives may be at work, which cannot be captured by payoffs alone.

Subjects in the dictator game may actually prefer the selfish outcome of keeping their entire endowment, but wish to maintain an illusion of not being selfish by choosing to act generously

[1]Some examples of distributional theories of social preference are altruism (Andreoni, 1990), which suggests that subjects may have an increasing utility in the payoffs of others, inequality aversion (Fehr and Schmidt, 1999), which suggest that subjects may be averse to advantageous inequality and utilitarianism and prioritarianism, which suggest that subjects may simply be attempting to maximise the sum total welfare of both players or maximising the payoff of the least well-off player, respectively.

regardless. DWK extend this idea further by suggesting that if this is truly the way in which subjects operate, then we could reasonably expect that some individuals will exploit opportunities which allow them to act in a selfish manner, so long as they are provided with a reasonable excuse to do so.

One such example of moral "wiggle room" which allows one to act selfishly without appearing so to others is the existence of uncertainty as to what exactly causes an unfair outcome. This concept known as "plausible deniability", suggests that subjects may exploit uncertainty as to what ultimately causes an unfair or inequitable outcome, in order to behave in a more self-interested manner. DWK investigate this phenomenon in an experimental setting by introducing a cut-off time to the decision of the dictator in the dictator game. After the cut-off point, if no decision has been made a computer will randomly make a choice on the subject's behalf. Crucially, only the dictator will ever know if a cut-off occurs, therefore, if the recipient in the dictator game observes an unequal final allocation, they have no way of knowing for certain, whether this outcome was explicitly chosen by the dictator or whether it was simply a product of chance.

DWK suggest that this uncertainty could provide subjects in the dictator game with the moral "wiggle room" in which to act in a more self-interested manner and therefore predict that the level of observed generosity in the dictator game will fall, when the cut-off feature is added. They find a rather striking result: of the 75% of subjects who weren't cut off by the computer, 55% chose the selfish outcome, compared to just 24% in the baseline dictator game.

One important question remains however: will subjects still exploit moral "wiggle room" in other experimental settings, such as the basic trust game, or is this effect simply a further artefact of the artificial, lab-based environment of the dictator game? This paper reports the results of an experiment which investigates whether trust fulfilling in the basic trust game is a desire for fair outcomes or a desire to maintain an appearance of fairness. More specifically it seeks to investigate whether or not the level of trust fulfilling decreases when plausible deniability is introduced via an act of chance, which chooses selfishly on the subject's behalf, with a given probability.

III. Experimental Design

The experimental design was the basic trust game as featured in Bacharach et. Al. (2007). This is a simultaneous move trust game in which both players face a set of binary choices: to "trust" or "withhold" trust for the first player (known as the truster) and to "fulfil" or "violate" trust for the second player (known as the trustee).

If the truster chooses to "withhold" their trust, both players earn £0. If the truster decides to trust, both players will receive £3 if the trustee then chooses to fulfil their trust. If the trustee chooses to violate trust, the truster will lose £3 but the trustee will gain £4.50.

Figure 1 – Basic Trust Game in Matrix Form

		Trustee	
		Fulfil	Violate
Truster	Trust	£3 £3	£4.50 -£3
	Withold	£0 £0	£0 £0

The strategy method was employed, meaning that all players in the role of the trustee were required to make their decision before knowing the decision of the truster, acting as if the truster had already chosen to trust them. This allows observations to be gathered for all subjects, even those whose decision as trustee was never taken into account, since the truster chose to withhold trust in the first stage. A mix of between and within-subjects design was also used. To increase the number of observations available for each of the two decisions, every subject made a decision as both the truster and the trustee, with no feedback given until both decisions were made. As the effects of plausible deniability on trust fulfilling is the main focus of this experiment, all subjects were required to make a decision first as the trustee in Task 1 and then immediately afterwards as the truster in Task 2.

All subjects were randomly assigned the role of truster or trustee but this role was only revealed to them after they had taken a decision in both roles. All subjects were randomly matched with another participant in the room who held the opposite role to them; this matching was strictly anonymous and no subject ever knew which other participant in the room that they had been matched with. Whilst the same subject acted in both roles within each treatment, a between-subjects design was used, meaning that each subject only ever faced one of two treatments.

The two experimental treatments are as follows: the baseline treatment which is simply the basic trust game as shown above and the plausible deniability treatment which has one additional feature. Now, when the subject makes a decision as the trustee as to whether or not to fulfil or violate trust, there is a 1 in 3 (33%) probability that an *act of chance* occurs which chooses to violate trust on their behalf. This probability was public knowledge. Whether or not an *act of chance* occurs was

determined by a private die role for each subject, once they had made their decision as the trustee in Task 1. If the number 1 or 2 was shown on the die, an *act of chance* occurred.

No subject could ever know whether or not an *act of chance* would occur and if it did occur, they were powerless to affect its choice of violate. Crucially, only the subject for whom the *act of* chance occurred would observe this occurrence. This ensured that the subject with which they were matched could not distinguish between the actions of the trustee or of the *act of chance* if their trust was violated, giving the trustee plausible deniability with which to violate the trust placed upon them. Note, however, that if the truster observed trust fulfilment, it was always a certainty that this was enacted by the trustee, since the *act of chance* would always choose to violate trust.

This *act of chance* plays a similar role to the cut-off feature in DWK. Rather than allowing for accidental cut-offs, which in the paper were assumed to be deliberate "self-deceptive" motives for not making an explicit choice, here all subjects are required to make a decision. Whilst subjects cannot guarantee a fair outcome, as in DWK, any subjects with a genuine preference for fair outcomes should continue to choose to fulfil trust, regardless of the presence of the *act of chance*.
Four sessions were run, two for each treatment, on the 18th March and the 24th March at the University of East Anglia. Subjects were a mixture of undergraduate and postgraduate students at the university and were recruited on a voluntary basis via university email (please see appendix A). All Instructions were given as a hard copy to each participant and read aloud by the experimenter, with short questionnaires being administered at suitable points to check understanding. A copy of the instructions and questionnaires can be found in Appendix B.

After the first two sessions were run, initial results suggested that there was no significant difference in the levels of trust fulfilling between session 1 and session 2 (χ^2=0.000, p=0.500 (2-tailed)) but that there was a significant increase in the number of subjects choosing to withhold their trust (χ^2=2.3449, p=0.063 (1-tailed)) . One interesting possibility was that the introduction of the *act of chance* may have altered the beliefs of subjects as to how other participants are likely to behave. With this in mind, session 3 and 4 included a brief belief elicitation task which took place immediately after the truster and trustee decision and before any feedback was given. Subjects were asked to guess, how many of the subjects in the room, including themselves, they thought had chosen to violate trust in Task 1 and how many had chosen to fulfil trust (See Appendix 3 for full instructions).

Payments were determined as follows. Subjects were given an initial credit of £4 to be used in Task 1 and Task 2, meaning all payoffs or losses from their decisions were in addition to this amount. After completing both tasks, each subject had a remaining credit which was the outcome of either Task 1 or Task 2, depending on their role. Each subject then privately drew one of two slips of paper marked "show-up fee" or "decision task" to determine whether they received their remaining credit from one of the decision tasks as payment or simply a £1 show-up fee. Additionally, subjects earned £1 from the belief task for guessing the exact proportion of subjects choosing to fulfil/violate trust and 50p for coming within 1 of the correct answer. Average earnings were £2.99, with a minimum of £1 and a maximum of £8.50. Each session lasted approximately 25 mins.

IV. Hypotheses

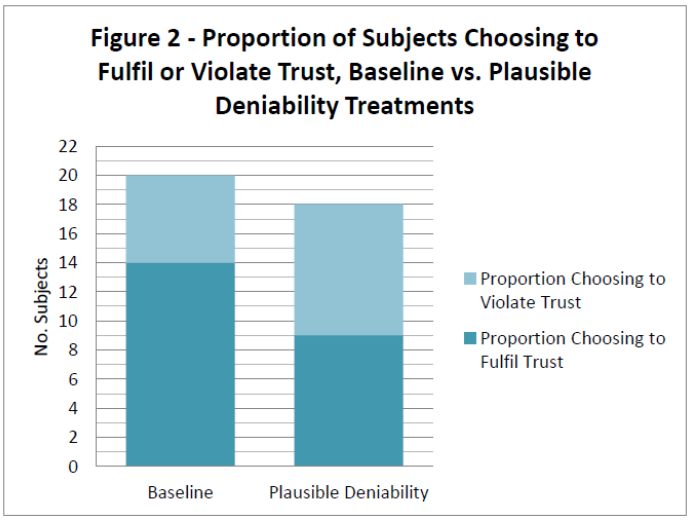
The null hypothesis would predict that there will be no difference in levels of trusting and trust-fulfilling between the baseline treatment and the plausible deniability treatment. The *act of chance* should have no effect on the levels of trust-fulfilling and should be irrelevant to subjects when making their decision. An alternative hypothesis, following the lines of DWK, would suggest that the *act of chance* will provide subjects with the moral "wiggle room" in which to act in a more self-interested manner and so the level of trust-fulfilling in the plausible deniability treatment should be lower than that in the baseline. Some subjects may choose to exploit uncertainty as to what exactly will cause a violation of trust in this setting, to violate trust more often themselves.

Any changes in the level of trusting should largely mirror those of trust-fulfilling, since any subject who chooses to violate trust in Task 1, can reasonably be expected to withhold their trust in Task 2. Another alternative could be that there is an effect on the level of trusting independent of changes to the level of trust-fulfilling. Trusters may recognise that the *act of chance* will provide plausible deniability for trustees and anticipate a higher level of trust violation, even if this ultimately does not occur. This would be demonstrated in the belief elicitation task by a difference in the stated beliefs of subjects in the baseline and plausible deniability treatment.

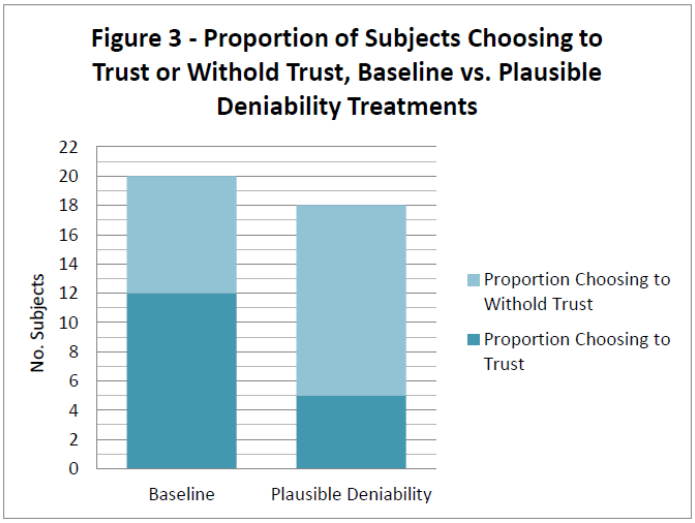
V. Results

The raw experimental data provide some idea as to which hypotheses are likely to be more convincing. As shown in figure 2, in the baseline treatment 14 of the 20 subjects (70%) chose to fulfil trust in Task 2, whilst the remaining 6 subjects (30%) chose to violate trust. In the plausible

deniability treatment, the number of subjects choosing to fulfil trust fell to 50%, with the remaining 50% of subjects choosing to violate trust.



The data for the decision of whether or not to trust or withhold trust is shown in figure 3. In the baseline treatment 12 of the 20 subjects (60%) chose to trust the person with which they were matched, whilst the remaining 8 subjects (40%) chose to withhold their trust. The level of trusting fell in the plausible deniability treatment to just 28%, with 13 of the 18 subjects (72%) choosing to withhold their trust.



Changes in trusting behaviour appear to mirror those of trust-fulfilling behaviour as would be expected. It is interesting to note however, that the average belief as to the proportion of subjects who would violate trust in Task 1 rose from 35% in the baseline treatment to 50% in the plausible

deniability treatment, possibly providing a reason for why the change in the level of trusting behaviour is much larger than that of trust-fulfilling behaviour. However, non-parametric techniques were used to test the significance of this difference and it was found that there was no significant change in the beliefs of subjects between the baseline and the plausible deniability treatment (Mann-Whitney, p=0.607).

Regression Results for Trust Fulfilling Behaviour

In addition to information as to the choices of subjects in Task 1 and Task 2, data was collected on each subject's gender, level of study, school of study and previous experience of participation in economics experiments at UEA. All of these categories were coded as dummy variables for the purpose of regression analysis. Gender took a value of 1 for males and 0 for females, level of study a value of 1 for postgraduate students and 0 for undergraduate students, school of study took a value of 1 for Economics and 0 otherwise and finally, previous experience took a value of 1 if the subject had past experience of participation in an Economics experiment and 0 otherwise.

As an initial precaution, a logit regression was run using the inclusion (or not) of the belief task as dummy variable (1=included, 0 otherwise), to confirm that the mere inclusion of the task in later sessions had no significant effect on subjects' behaviour. This regression found that the belief task itself had no significant effect on the level of trust-fulfilling (p=0.137).

To empirically test the effects of introducing plausible deniability into the basic trust game, a probit regression was run in STATA of the decision to fulfil or violate trust on the following dummy variables: treatment (1=Plausible Deniability, 0=Baseline) , gender, level of study, economics or non-economics, previous experience and an interaction term between economics and previous experience.

Table 1 - Results of the Probit Regression for the Trustee Decision

Variable	Expected Sign	Coeff.	Std. Err.	Z	P>|z|	95% Conf. Int.	
Treatment**	-	-1.1416	0.5836	-1.96	0.050	-2.2855	0.0023
Gender***	-	-1.6369	0.5985	-2.74	0.006	-2.8098	-0.4639
Ugpg	?	0.8192	0.6680	1.23	0.220	-0.4901	2.1285
Economics	-	-1.8551	1.1418	-1.62	0.104	-4.0930	0.3828
Experience**	-	-2.7843	1.2868	-2.16	0.030	-5.3064	-0.2622
EconExp**	-	-2.8821	1.4684	-1.96	0.050	0.0041	5.7601
Constant***		2.7203	1.0161	2.68	0.007	0.7286	4.7119
No. Observations =	38						
LR Chi2 =	15.34						
Prob > Chi2 =	0.0177						
Pseudo R2 =	0.3010						
Log Likelihood =	-17.8187						

As can be seen from Table 1, when plausible deniability is added to the basic trust game, subjects are less likely to fulfil trust than those subjects in the baseline treatment, all other things being equal (p=0.050). The effect of gender on the decision of whether or not to fulfil trust is highly significant (p=0.006). Males are less likely to fulfil trust than females, ceteris paribus. The effect of degree level was found to be insignificant on the decision of whether or not to fulfil trust (p=0.220).

There is a significant joint effect of economics and experience (given by the interaction term EconExp) on the decision of whether or not to fulfil trust (p=0.050). This suggests that Economics students with previous experience of Economics experiments are less likely to fulfil trust, holding all other factors constant.

The non-parametric counterparts to the above probit regression were also carried out, but no significant difference was found between the choices of subjects in the baseline treatment and those in the plausible deniability treatment (χ^2=1.586, p=0.208). The only significant effect was gender (χ^2=4.716, p=0.030). Whilst degree level was insignificant as in the regression above (χ^2=0.035, p=0.851) the chi-square test also suggested that there was no significant difference in the choices of Economics students with previous experience of Economics experiments (represented by the interaction term in the regression) compared to all other subjects (χ^2=0.741, p=0.389) .

Regression Results for Trusting Behaviour

The test the effects of introducing plausible deniability in the basic trust game on the level of trusting, a probit regression was run in STATA of the decision to trust or 8ithhold trust on the following dummy variables: treatment (1=Plausible Deniability, 0=Baseline), gender, level of study, economics or non-economics, previous experience and an interaction term between gender and economics. Once again, an initial regression was run to confirm that the belief task itself had no significant effect on the levels of trusting (p=0.897).

As can be seen from Table 2, when plausible deniability is added to the basic trust game, subjects are less likely to trust others than subjects in the baseline treatment, all other things being equal (p=0.042). The effect of degree level was found to be insignificant on the decision of whether or not to trust or withhold trust (p=0.117). Previous experience of Economics experiments appears to have a significant and positive effect on trust (p=0.050); if a subject has previously participated in

Economics experiments, they are more likely to trust than if they have no prior experience, ceteris paribus.

Table 2 - Results of the Probit Regression for the Truster Decision

Variable	Expected Sign	Coeff.	Std. Err.	z	P>|z|	95% Conf. Int.	
Treatment**	-	-1.2554	0.6187	-2.03	0.042	-2.4680	-0.0428
Gender**	-	2.4884	1.2360	2.01	0.044	0.0658	4.9109
Ugpg	?	1.2501	0.7977	1.57	0.117	-0.3134	2.8135
Economics	-	0.8932	0.7544	1.18	0.236	-0.5854	2.3718
Experience**	-	1.2069	0.6145	1.96	0.050	0.0025	2.4114
GenEcon**	-	-4.1108	1.6732	-2.46	0.014	-7.3902	-0.8314
Constant*		-1.6076	0.9648	-1.67	0.096	-3.4985	0.2833
No. Observations =	38						
LR Chi2 =	16.17						
Prob > Chi2 =	0.0129						
Pseudo R2 =	0.3093						
Log Likelihood =	-18.0461						

There is a significant joint effect of economics and gender (given by the interaction term GenEcon) on the decision of whether or not to fulfil trust (p=0.014). This suggests that male Economics students are more likely to withhold trust than other students, holding all other factors constant.

The non-parametric counterparts to the above probit regression were also carried out. A significant difference was found between the choices of subjects in the baseline treatment and those in the plausible deniability treatment (χ^2=3.979, p=0.046). The only other significant difference was found with the choices of male Economics students (represented by the interaction term in the regression) compared to all other subjects (χ^2=2.763, p=0.096). Degree level was insignificant as in the regression above (χ^2=1.304, p=0.254) but the chi-square test also suggested that there was no significant difference in the choices of students with previous experience of Economics experiments and those with no prior experience (χ^2=2.343, p=0.126).

VI. Discussion

It appears from the above regressions that there does seem to be an effect on both the levels of trust-fulfilling and trusting of incorporating plausible deniability into the basic trust game. It's difficult to say whether or not these results would hold in a study conducted under different conditions, such as a larger sample size, and there is every possibility that the difference between the two treatments could become either more or less significant in future experiments, but so far the results look promising.

The differences in the level of trust-fulfilling could be explained by the moral "wiggle room" motive suggested by DWK. Subjects may consider the *act of chance* to be providing sufficient moral "wiggle room" within which to act in a more selfish manner. The fact that there is uncertainty as to what causes a violation of trust in the plausible deniability treatment, could be providing the perfect excuse for those subjects who would prefer to violate trust but would also like to maintain an appearance of fairness towards others.

It's worth noting some of the limitations of this experiment, the first of which is that the belief task was only administered to a small portion of the sample size. This may be why the difference in beliefs between the two treatments was found to be insignificant. With a larger sample size, we could reasonably expect that a change in trusting and trust-fulfilling behaviour would be mirrored by a change in the beliefs of subjects; either subjects would update their beliefs based on the actions which they themselves took in the task, or the introduction of the *act of chance* itself would alter the beliefs of subjects which, in itself, could lead to part of the observed changes in the level of trusting and trust-fulfilling.

Secondly, the average payments for this experiment were quite low by normal standards and may have lead to unusual or uncharacteristic behaviour. There is also a chance that the dual-role approach to subject decisions may have left some subjects confused as to their interactions with other players in the room. Whilst no subjects expressed any confusion or lack of understanding during the instructions, questionnaires or demographic questionnaire after the experiment, some subjects' stated methods of choosing between the two options in Task 1 seem to convey a belief that the decision they make in the first task will affect the earnings that they will receive in the second task, despite that fact that it is only their partner's decision which matters. With a larger budget, subjects would only need to take on one role to reduce possible confusion.

VII. Further Research

The results from this experiment are promising and suggest that further research should be made into the effects of moral "wiggle room" in domains outside of the artificial setting of the dictator game. Whilst further research may prove that this effect is largely diminished or completely wiped out with a larger sample size, this pilot study would suggest that moral "wiggle room" has the potential to be robust to other motives which are traditionally at work in the trust game setting, such as reciprocity. This would in itself present a rather important finding: even in situations where

making a decision with money which has been provided by another trusting individual, some people could still prefer to violate the trust placed upon them, but choose to fulfil trust regardless, out of a preference to maintain an illusion of fairness towards others. If given the plausible deniability with which to act in selfish manner, some individuals may choose to violate trust which is placed upon them.

VIII. Conclusion

This experiment investigated the effects on trust-fulfilling, of introducing "plausible deniability" into the basic trust game via an *act of chance* which violates trust on a subject's behalf with some given probability. The initial results from the probit regression which was run suggest that there is a significant decrease in the level of both trusting and trust-fulfilling, when plausible deniability is introduced into the trust game. This would suggest that some individuals may actually prefer to violate trust, but are choosing to fulfil trust out of a desire to maintain an appearance of fairness to others. When given the plausible deniability within which to act in a self-interested manner, some individuals choose to violate trust and exploit the uncertainty as to what exactly causes a violation of trust. Whilst these initial results are promising, further study using a larger sample size, a single-role for participants and larger monetary incentives would need to be conducted to account for the limitations of the current experiment and to test the robustness of the results which have been presented.

References

Andreoni, J. (1990) “Impure Altruism and Donations to Public Goods: A Theory of Warm Glow Giving” *Economic Journal, pp. 464-477.*

Bacharach, M., Guerra, G. & Zizzo, D. J. (2007) “The Self-Fulfilling Property of Trust: An Experimental Study” *Theory and Decision, pp. 349-388.*

Camerer, C. (2003) “Behavioural Game Theory: Experiments on Strategic Interaction” *Princeton University Press: Princeton, NJ.*

Dana, J., Weber, R. A. & Kuang, J. X. (2007) “Exploiting Moral Wiggle Room: Experiments Demonstrating an Illusory Preference for Fairness” *Economic Theory, pp. 67-80.*

Fehr, E. & Schmidt, K. M. (1999) “A Theory of Fairness, Competition and Cooperation” *Quarterly Journal of Economics, pp. 817-868.*