"They do what I do": Positive Correlation in Ex-Post Beliefs

Ariel Rubinstein
Tel Aviv University and New York University
and
Yuval Salant
Northwestern University

Abstract. After playing the Chicken game, players report their beliefs about their opponent's strategy in different frames. It is found that the framing of the belief elicitation question influences the degree of positive correlation between a player's reported belief and his action. It is also found that players believe there is positive correlation in the actions of two randomly selected players. The results are consistent with the existence of two forces influencing ex-post belief formation: strategic justification (i.e. a player wishes to rationalize his action in the game) and self-similarity (i.e. a player thinks that his opponent behaves the same way he does). Implications for belief elicitation techniques and equilibrium concepts are discussed.

The first author acknowledges financial support from ERC grant 269143.

We thank Hadar Binsky for his assistance in analyzing the data, and Eli Zvuluny for constructing and maintaining the web platform on which the research is conducted. We also thank Karl Schlag, Andy Schotter, Jörg Spenkuch, and James Tremewan for their feedback.

1. Introduction

This paper studies players' ex-post beliefs about their opponent's strategy in a game. Ex-post beliefs are of interest because they may reflect players' ex-ante beliefs. They may thus assist in determining whether players best-respond to what they believe to be their opponent's strategy. It is a common practice in experimental game theory to refrain from asking players about their ex-ante beliefs because this may affect their deliberation process and actions in the game (for a different view, see Costa-Gomes and Weizsäcker (2008)).

We hypothesize that ex-post beliefs in strategic interactions are influenced by two forces:

(1) **"Strategic justification"**: a player who is asked ex-post about his beliefs will have a tendency to report beliefs that rationalize his action in the game whether or not they actually influenced his action.

(2) **"Self-similarity"**: a player tends to think that other players behave similarly to him, and thus will report beliefs that are biased toward his own action.

Our main experiment uses the Chicken game to demonstrate the existence of the two forces. In this game, strategic justification and self-similarity operate in opposite directions because rationalizing any action in the game requires assigning a high probability to your opponent choosing a different action. Participants' reported beliefs may therefore differ when they are elicited in a frame that aims to trigger the activation of strategic justification and a frame that aims to trigger the activation of self-similarity.

Participants in the experiment were told that they are about to play the following game against an opponent selected randomly from among several hundred participants:

Your choice	Your opponent's choice	
	Dove	Hawk
Dove	30, 30	20, 70
Hawk	70, 20	0, 0

After choosing their action in the game, they were asked about their ex-post beliefs in one of the following three frames:

(i) The "opponent" frame – The participant was asked to state his beliefs regarding the choice of his randomly selected opponent.

(ii) The "population" frame – The participant was asked to state his beliefs regarding the distribution of choices among all the participants.

(iii) The "outcome" frame – The participant was asked what he believes to be the distribution of outcomes (i.e., pairs of actions) among all the pairs of participants.

Given that a player's opponent is randomly chosen from among several hundred participants, identical beliefs should be reported in the opponent frame and the population frame. Furthermore, responses in the outcome frame should be consistent with the reported beliefs

in other two frames and the assumption that players' actions are independent.

Participants' reported beliefs, however, place significantly more weight on a participant's own action in the population frame than in the opponent frame. This is consistent with the hypothesis that the opponent frame triggers strategic justification whereas the population frame triggers self-similarity.

There are two main findings in the outcome frame. First, participants tend to place the largest weight on the outcome in which the two randomly chosen players take the same action as they do. That is, the outcome frame invokes self-similarity, as in the case of the population frame. Second, participants express a positive correlation between the actions of two randomly chosen players. Specifically, participants assign significantly more weight to outcomes in which the two players choose the same action than is warranted by the independence assumption.

The finding that players' beliefs violate the independence assumption may motivate the development of a new game-theoretic solution concept that takes this systematic failure into account. One possible approach is to assume that each player forms correct beliefs on the actions of every other player, but when he assigns probabilities to action profiles, he distorts the probabilities in a way that assigns higher probability than is consistent with the independence assumption to profiles in which players take similar actions. We propose such a solution concept in the Discussion section.

Our findings are also relevant for the assessment of belief elicitation techniques in experimental game theory. Those methods were recently surveyed by Schlag, Tremewan and van der Weele (2014) and Schotter and Trevino (2014). The differences in reported beliefs between the opponent and population frames imply that players' elicited beliefs may depend on the framing of the belief elicitation question. Furthermore, the failure of independence implies that eliciting players' beliefs by asking about other players' actions may yield different results than eliciting them by asking about outcomes.

Related literature:

(a) Self-similarity in non-strategic settings. Ross, Greene & House (1977) report a series of studies illustrating that people tend to perceive their own choices and judgements as being relatively common in the population. This False Consensus effect has been documented since then in a variety of settings, and explained by various psychological theories (see Marks and Miller (1987) for a survey). In fact, the effect is so strong that it appears even in context-free environments. To illustrate this, we asked several hundred participants (from a population similar to the one used in the main experiment) to pick an integer between 1 and 9, and then to estimate the distribution of answers among all participants. A monetary reward was promised to the participant whose prediction is closest to the actual distribution of responses. After omitting the answers of participants whose reported be-

liefs did not sum up to 99% or 100%, the distribution of the choices of the 546 remaining participants was (7%, 5%, 11%, 8%, 11%, 8%, 24%, 18%, 9%). In line with the False Consensus effect, participants' average estimation of the frequency of their own choice in the population was 15% rather than 11% as expected. Moreover, the average estimate for each number by those that chose it was strictly larger than the estimate for the same number among those that did not.

(b) The experimental literature on beliefs in the Prisoner's Dilemma (PD) game. Dawes, McTavish and Shaklee (1977) and Messé and Sivacek (1979) find that in the PD game players tend to attribute their own action to other players, regardless of which action they chose. Dawes, Robyn, and Shaklee (1977) conjecture that either strategic justification or self-similarity or both may explain these findings. We repeated the main experiment for the PD game. In the Discussion section, we report the results and discuss the connections to this literature.

2. Procedure

The platform for the experiment was the didactic site http://gametheory.tau.ac.il. The participants were students from 40 countries who had taken a course in game theory. They had used the site previously and agreed to participate in additional online survey experiments.

Participants were told that they were about to play a game against an opponent who would be selected randomly from a population of several hundred participants. They were informed that one pair of players would be chosen randomly from among all pairs playing the game, and would receive a payment (in USD) according to the outcome of their game.

After choosing their action in the game, each participant was randomly assigned to one of the following three frames, in which they were asked about their beliefs:

(i) The "Opponent" frame: What are your beliefs regarding the choice of your opponent?

I assign a probability of % to my opponent choosing Dove and % to my opponent choosing Hawk.

(ii) The "Population" frame: What are your beliefs regarding the distribution of choices among all those who play the game?

I believe that % choose Dove and % choose Hawk.

(iii) The "Outcome" frame: What are your beliefs regarding the distribution of outcomes for all pairs who play the game?

I believe that in % of the games, both players choose Dove.

I believe that in % of the games, both players choose Hawk.

I believe that in % of the games, one player chooses Dove and the other chooses Hawk.

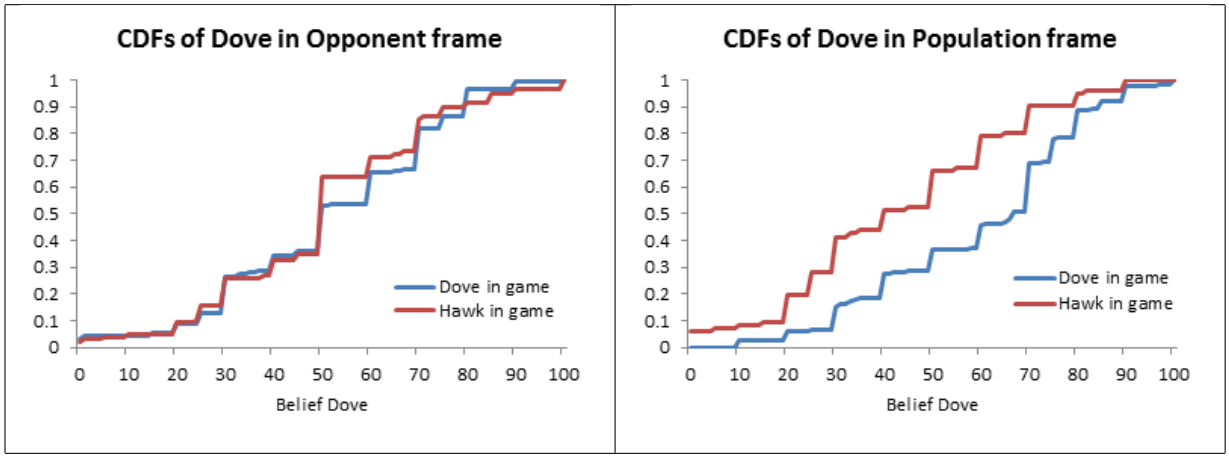
Subjects whose answers did not sum up to 100% were removed from the data leaving 718 participants.

3. Results

About 62% of the 718 participants chose Dove and 38% chose Hawk. The following table presents the average reported beliefs in the opponent and the population frames depending on whether Dove or Hawk was chosen in the game.

	Opponent frame		Population frame	
Action in game	Dove (N=142)	Hawk (N=97)	Dove (N=162)	Hawk (N=82)
Average belief Dove	51.8% (1.9)	50.5% (2.3)	59.9% (1.7)	43.3% (2.6)
Average belief Hawk	48.2%	49.5%	40.1%	56.7%

The following two diagrams present, for every action and every frame, the CDF of the probability assigned to Dove conditional on the action and the frame. That is, given an action and a frame, we plot for every $0 \leq x \leq 100$ the proportion of participants who assigned a belief weakly smaller than x to Dove from among those who chose the given action in the given frame.



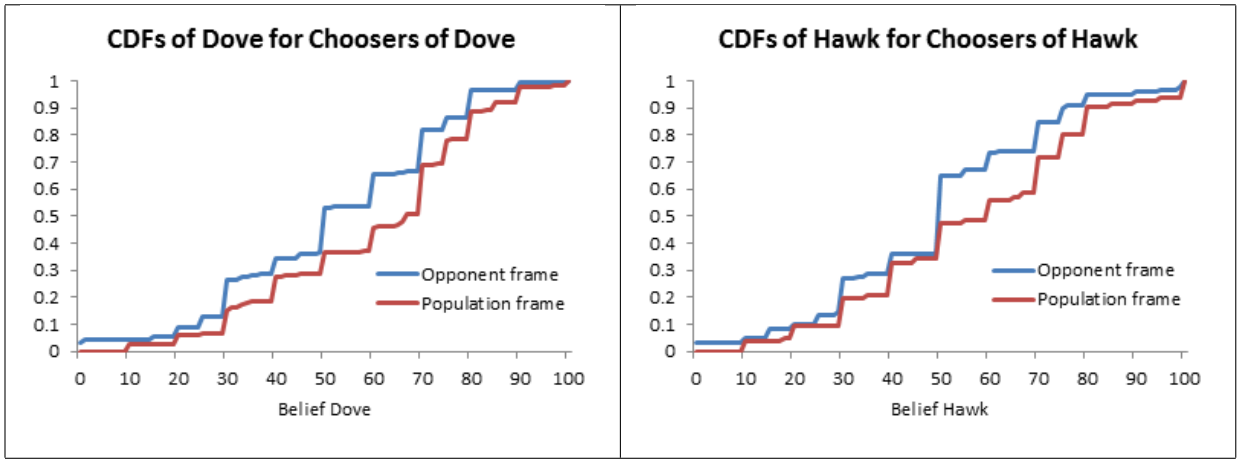
The CDFs in the opponent frame are very similar to one another.[1] Thus, if participants maximize expected utility given their beliefs, it must be that the attitude toward risk of those who chose Dove (and essentially faced little uncertainty) differs from that of those who chose Hawk (and faced significant uncertainty).

The CDFs in the population frame are very different from one another.[2] In the diagrams, the CDF of the probability assigned to Dove by choosers of Dove first-order stochastically dominates the CDF of the probability assigned to Dove by choosers of Hawk.

[1] The Kolmogorov–Smirnov test statistic for the equality of the CDFs is 0.11 ($p = 0.45$).

[2] The K–S test statistic for the equality of the CDFs is 0.35 ($p < 0.001$).

The next two diagrams provide another illustration. The left diagram compares the CDFs of Dove in the two frames for choosers of Dove, and the right diagram compares the CDFs of Hawk in the two frames for choosers of Hawk. In the diagrams, participants' beliefs place more weight on their own action in the population frame than in the opponent frame.[3]



This finding is consistent with the hypothesis that the population frame leads players to put more weight on self-similarity in forming ex-post beliefs, whereas the opponent frame leads players to put more weight on strategic justification.

We turn to the outcome frame. The following table presents participants' average reported beliefs in this frame according to the action chosen in the game:

	Belief in Outcome frame	
Choice in game	Dove (N=142)	Hawk (N=93)
(Dove, Dove)	45.6% (2.1)	28.3% (2.3)
(Hawk, Hawk)	28.7% (1.7)	42.1% (3.0)
(Dove, Hawk)	25.7% (1.3)	29.6% (2.4)

The table provides additional evidence that there is a bias among choosers of a particular action towards believing that others choose the same action. In order to estimate the bias, we infer the marginal beliefs of players over actions from their reported beliefs over outcomes. Denote by d the probability that a given participant assigns to the two players choosing Dove and by h the probability that he assigns to the two players choosing Hawk. Thus, he assigns a probability of $(1-d-h)$ to the event that one player chooses Dove and the other chooses Hawk. This implies that he believes that the frequency of players who chose Dove in the population is $d+\frac{1-d-h}{2}$ and the frequency of those who chose Hawk is $h+\frac{1-d-h}{2}$.

[3]The K–S test statistic for the equality of the CDFs in the left diagram is 0.2 ($p<0.005$) and in the right diagram is 0.18 ($p=0.09$).

Using this method to estimate the induced beliefs over actions from the reported beliefs over outcomes, we find that on average participants who play Dove believe that 58.5% of the other players play Dove as well, while participants who play Hawk believe that only 43.1% play Dove. These proportions are very close to the average reported beliefs in the population frame, suggesting that the outcome frame activates self-similarity in a similar way to the population frame.

Another finding in the outcome frame is that players believe there is positive correlation between the actions of two randomly chosen players in the sense that they place more weight on the two players choosing the same action than is warranted by the independence assumption. To demonstrate this, define the excess probability that a participant assigns to the correlation between two randomly chosen players by $\Delta_1 = [d + h] - [(d + (1 - d - h)/2)^2 + (h + (1 - d - h)/2)^2]$. That is, we measure the distance between the actual weight a participant assigns to the two players choosing the same action and the weight implied by the participant's induced marginal beliefs and the independence assumption.

Another possible metric is $\Delta_2 = \sqrt{d} + \sqrt{h} - 1$, which measures how far the vector $(\sqrt{d}, \sqrt{h})$ is from being a belief. (To illustrate: if $d = 0.6$ and $h = 0.3$, then $\Delta_1 = 0.355$ and $\Delta_2 = 0.322$). In both metrics, the reported beliefs are consistent with independence if and only if $\Delta_i = 0$.

The following table demonstrates that a vast majority of participants expressed a positive correlation between the actions of two randomly chosen players, while only 15% satisfied the independence assumption:

Sign of Δ	% ($N = 235$)	Average Δ_1	Average Δ_2
> 0	74.0%	18.5% (0.7)	18.8% (0.6)
0	14.9%	0	0
< 0	11.1%	−17.6% (3.4)	−30.5% (6.6)

4. Discussion

A. Modelling the failure of independence. One may argue that the failure of independence is an indication that people believe in the existence of a convention dictating how to play the game, but are uncertain as to what that convention is. To illustrate this argument, suppose that there is a convention in society to either "Play Dove" or "Play Hawk", and that the two are initially equally likely. Each player obtains a signal to "Play Dove" or to "Play Hawk" that is identical to the convention with probability $x > 1/2$, and differs from it with probability $1 - x$. The player acts according to his signal. Bayesian updating then implies that conditional on obtaining the signal "Play Dove", $d = x^3 + (1 - x)^3$ and $h = x(1 - x)^2 + (1 - x)x^2$. The assumption that $x > 1/2$ is equivalent to the statement that $\Delta_i > 0$.

Such a model does not explain the experimental results. This is because the signal has to be very informative in order to place enough weight on (Dove, Dove) and (Hawk, Hawk). But such an informative signal implies that a player should place a much larger weight than in the experimental results on the outcome in which the two randomly chosen players choose the same action that he did. In particular, obtaining $d + h$ in the vicinity of 75% (as in the experiment) requires the signal's accuracy to be about 0.85. However, at this level of accuracy, the ratio d/h conditional on choosing, say Dove, is around 4.8 which is nowhere near the experimental results.

B. Alternative equilibrium notion. The failure of independence may motivate the development of a solution concept in which a player's beliefs about other players' actions would be allowed to violate the independence assumption. Consider, for example, a setting in which three firms are deciding whether to enter a new market. The profit of a firm that decides to enter is G if the other two firms do not, 0 if only one other firm enters, and $-B$ if both of the other two enter, where $B, G > 0$. The standard Nash equilibrium conditions are that each firm makes an optimal decision given its beliefs over the strategy profiles of the two other firms, where these beliefs are correct in the sense that (a) marginal beliefs over the actions of any other firm are consistent with this firm's strategy and (b) beliefs over strategy profiles are derived from the marginal beliefs using the independence assumption.

In the unique symmetric Nash equilibrium of this game, the entry/no entry ratio is $\sqrt{G}/\sqrt{B}$ (denoting the probability of entry by x, the equilibrium has to satisfy $G(1-x)^2 - Bx^2 = 0$).

Assumption (b) in the definition above can be changed in order to account for a systematic failure of independence. For example, suppose that each firm has the correct marginal beliefs but expects perfect correlation between the actions of the other two firms. The new equilibrium entry probability x then has to satisfy $G(1-x) - Bx = 0$, which leads to an entry/no entry ratio of G/B. Thus, replacing the independence assumption with perfect correlation increases the probability of entry if $G > B$, and decreases it if $G < B$.

C. The Prisoner's Dilemma game. As mentioned in the Introduction, the experimental literature on the Prisoner's Dilemma (PD) game has already documented that players in the PD game tend to attribute their own action to others playing the game. Dawes, McTavish and Shaklee (1977) also conjectured that either strategic justification or self-similarity (or both) may explain this finding in the PD game.

We find the Chicken game to be more appropriate than the PD game in order to demonstrate the presence of strategic justification and self-similarity. The problem in the PD game is that if players view their payoffs as utilities, then it is impossible to

strategically justify cooperation. And if, as Dawes, McTavish and Shaklee (1977) argue, players strategically justify cooperation by believing that their opponent will cooperate, then self-similarity and strategic justification operate in the same direction in this game. Thus, unlike in the Chicken game, the reported beliefs in the PD game should be similar regardless of how the belief elicitation question is framed.

To test this, participants who had played the Chicken game, were then asked to repeat the same procedure for the PD game with the following payoff matrix:

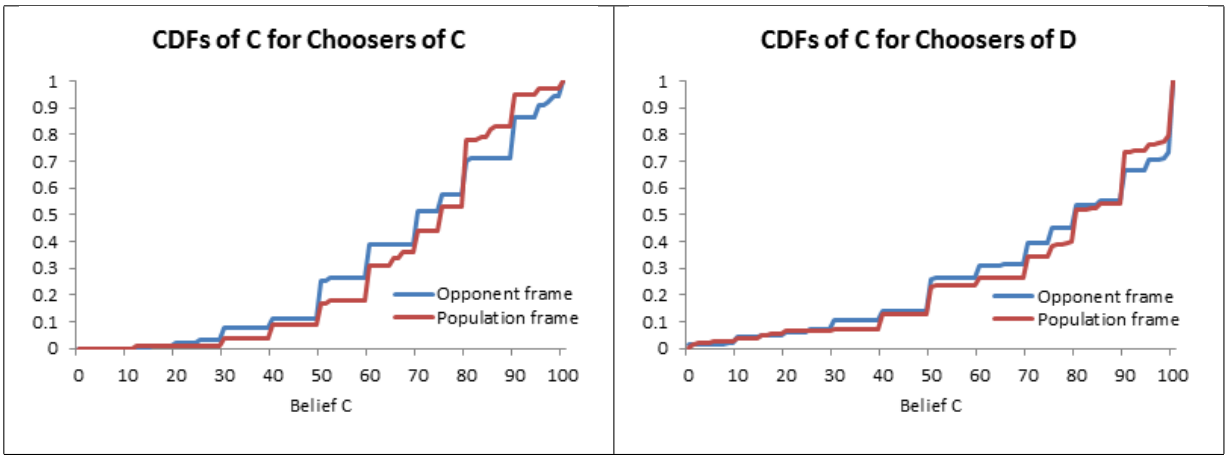
Your choice	Your Opponent's choice	
	C	D
C	50,50	0,60
D	60, 0	10,10

About one-third of the 718 participants chose to cooperate. The following two tables present the average reported beliefs in the opponent and population frames depending on whether participants chose to cooperate (C) or defect (D).

	Opponent frame		Population frame	
Action in game	C (N=90)	D (N=167)	C (N=77)	D (N=147)
Average belief C	69.7% (2.2)	26.1% (2.1)	70.8% (2.0)	25.0% (2.1)
Average belief D	30.3%	73.9%	29.2%	75.0%

It is clear from the tables that there is a strong tendency among participants to attribute their own action to other players. This is very much in line with the results in Dawes, McTavish and Shaklee (1977) and Messé and Sivacek (1979). However, this tendency is not affected by the frame, i.e., there is no difference in average reported beliefs between the two frames. There is also no difference in the distributions of reported beliefs, as can be seen in the following two diagrams. The left diagram compares the CDFs of C for choosers of C in the two frames, and the right diagram compares the CDFs of D for choosers of D.[4] These results are consistent with the hypothesis that strategic justification and self-similarity operate in the same direction in the PD game.

[4]The K–S test statistic for the equality of the CDFs in the left diagram is 0.12 ($p = 0.56$) and in the right diagram is 0.08 ($p = 0.73$).



Messé and Sivacek (1979) conducted a related experiment in which they asked participants to play the PD game, and report their beliefs about either their opponent or a participant in another dyad. In contrast to the findings presented here, they find that subjects who were asked about a player in another dyad believed this player would take the same action as they did to a lesser extent than when asked about their own opponent.

The findings in the outcome frame are quite similar to those in the Chicken game. First, there is a bias among choosers of a particular action towards believing that others choose the same action, as illustrated in the following table:

	Belief in outcome frame	
Choice	C (N=71)	D (N=166)
(C,C)	63.8% (2.8)	18.2% (1.7)
(D,D)	17.5% (1.8)	63.0% (2.3)
(C,D)	18.7% (1.8)	18.8% (1.5)

,

Second, a large majority of the participants believe there is positive correlation between the actions of two randomly chosen players:

Sign of Δ	% ($N = 237$)	Average Δ_1	Average Δ_2
> 0	70.0%	17.0% (0.7)	19.2% (0.6)
0	17.3%	0	0
< 0	12.7%	-9.7% (2.5)	-17.0% (4.5)

Thus, the tendency to to express positive correlation between the actions of two randomly chosen players extends to the PD game.

D. Belief elicitation and incentives. The belief elicitation procedure used in this paper was simply to ask participants ex-post about their beliefs. A popular procedure in the experimental literature in economics incentivizes participants to report truthfully

by providing them with a probabilistic reward scheme with the following property: given the participant's beliefs, the expected value of the reward is maximized if the participant truthfully reports his beliefs. The Quadratic Scoring Rule method is probably the most commonly used incentive-based elicitation method. It elicits participants' beliefs regarding an event A by asking them to choose a number x between 0 and 100, and rewards them with $C-\frac{(100-x)^2}{100}$ if A occurs and $C-\frac{(0-x)^2}{100}$ if it does not. That is, participants are penalized according to the square of the distance between the specified x and the actual outcome. If a participant assigns the probability p to the event A and if he maximizes expected payoff, (i.e.,$C+p[100-(100-x)^2/100]+(1-p)[100-(-x)^2/100]$), then he should report $x=100p$, i.e., his beliefs about the event A.

In order to test whether participants are indeed able to understand and solve this problem, we asked students who had completed a game theory course to choose a number x between 0 and 100 and to imagine that they will be paid $100-(V-x)^2/100$, where V is a random variable. Half of the participants were told that V receives the value 100 with probability 30% and 0 with probability 70%. The other half were told that V receives the value 100 with probability 70% and 0 with probability 30%. That is, we explicitly specified the correct belief. The distributions of responses and median response times (for the categories with a significant number of participants) are reported below:

Values	0	30	50	70	100	Other	N	Average
$E(V)=30$	33%	26%	13%	4%	10%	15%	392	32.1%
MRT	130	204	125		111			
$E(V)=70$	15%	3%	13%	25%	26%	18%	413	58.6%
MRT	94		132	167	104			

The results of the two versions are remarkably similar. Only one quarter of the participants chose the correct belief and as expected (see Rubinstein (2013)) they spent substantially more time solving the problem than the rest. It is also interesting that the average belief of all subjects is close to the correct belief. However, it should also be noted that even in the top quartile of participants in terms of response time, only 39% reported the correct belief.

These results may hint that incentive-based methods for eliciting beliefs may not always perform better than simply asking the participants what their beliefs are.

References

Costa-Gomes, Miguel A. and Georg Weizsäcker (2008). "Stated beliefs and play in normal-form games". *The Review of Economic Studies*, 75(3), 729-762.

Dawes, Robyn M., Jeanne McTavish, and Harriet Shaklee (1977). "Behavior, communication, and assumptions about other poeple's behavior in a commons dilemma situation". *Journal of Personality and Social Psychology*, 35(1), 1–11.

Marks, Gary and Norman Miller (1987). "Ten years of research on the false-consensus effect: An empirical and theoretical review". *Psychological Bulletin*, 102 (1), 72-90.

Messé, Lawrence and John M. Sivacek (1979). "Predictions of others' responses in a mixed-motive game: self-justification or false consensus?". *Journal of Personality and Social Psychology*, 37(4), 602–607.

Ross, Lee, David Greene, and Pamela House (1977). "The "false consensus effect": An egocentric bias in social perception and attribution processes". *Journal of Experimental Social Psychology*, 13(3), 279-301.

Rubinstein, Ariel (2013). "Response Time and Decision Making: An Experimental Study". *Judgement and Decision Making*, 8(5), 540-551.

Schlag, Karl, James Tremewan, and Joel van der Weele (2014). "A Penny for Your Thoughts: A Survey of Methods for Eliciting Beliefs." University of Vienna, Department of Economics.

Schotter, Andrew and Isabel Trevino (2014). "Belief Elicitation in the Lab". *Annual Review of Economics*, Vol. 6.