CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The University of Nottingham

Discussion Paper No. 2011-08

Alessia Isopi,
Daniele Nosenzo
and Chris Starmer
July 2011

Does Consultation Improve Decision Making?

CeDEx Discussion Paper Series
ISSN 1749 - 3293

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Sue Berry
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 15469
Fax: +44 (0) 115 95 14159
sue.berry@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers

Does consultation improve decision making?

by

Alessia Isopi*, Daniele Nosenzo* and Chris Starmer*

22 July 2011

Abstract

This paper reports an experiment designed to test whether prior consultation within a group affects subsequent individual decision making in tasks where demonstrability of correct solutions is low. In our experiment subjects considered two paintings created by two different artists and were asked to guess which artist made each painting. We observed answers given by individuals under two treatments: in one, subjects were allowed the opportunity to consult with other participants before making their private decisions; in the other there was no such opportunity. Our primary findings are that subjects in the first treatment evaluate the opportunity to consult positively but they perform significantly worse and earn significantly less.

Keywords: Consultation; Decision making; Group decisions; Individual decisions.

JEL Classification Numbers: C91; C92; D80.

Acknowledgements: We thank Spiros Bougheas, Niall Flynn, Jeroen Nieboer and Martin Sefton for helpful comments. Alessia Isopi acknowledges support from the Marie Curie Foundation (project TAAWE/220489). Daniele Nosenzo acknowledges support from the Leverhulme Foundation (ECF/2010/0636).

* School of Economics, University of Nottingham, Nottingham, NG7 2RD, United Kingdom.
Corresponding author: Chris Starmer, email: chris.starmer@nottingham.ac.uk.

1. INTRODUCTION

In many walks of life, individuals often consult, either formally or informally, with others before taking important decisions. Obvious examples include borrowers or investors consulting with financial advisors before making decisions; or employers consulting with hiring managers and HR specialists before hiring new employees; or individuals talking with family or friends, in addition to health professionals, in advance of choosing between alternative medical interventions. With the growth of the internet, sources of 'advice' are expanding rapidly while the costs of accessing them are often very low. But here, as elsewhere, the quality of advice obtained may be difficult to assess raising interesting questions about when advice should be followed and the conditions under which consulting with others can be expected to improve (or worsen) individual decision making.

One body of literature which might inform understanding of the influence of consultation on individual decisions is an extensive literature that has examined the comparative success of decisions made by individuals versus decisions made by groups. A considerable body of evidence now supports the claim that groups can often 'outperform' individuals. The bulk of this literature comes from experimental research in social psychology examining behavior in decision problems that have correct solutions and thus have a meaningful criterion for assessing decision accuracy. Within this literature, a widely reported finding is that groups are more likely to report the correct answer (see, e.g., Hastie, 1986; Laughlin et al., 2003; Laughlin et al., 2006, and references therein). Economists have also compared individual and group decisions with most of this research focused on interactive decisions[1] where the players are either groups of individuals or single individuals; in this context, a common result, across a variety of games, is that groups' decisions more closely track standard game theoretic predictions.[2]

[1] There is a smaller literature investigating the incidence of preference 'anomalies' comparing groups and individuals. For example, Charness et al. (2007) find that failures to perform Bayesian updating and violations of first-order stochastic dominance decrease with the size of the group. Sutter (2007) finds that, whereas groups are as prone to as individuals to myopic loss aversion, group decision-making attenuates its effects. Some studies report that groups are as prone as individuals to other decision anomalies, e.g. the endowment and the compromise effect (Munro and Popov, 2009), and violations of expected utility theory (Bone et al., 1999; Rockenbach et al., 2007).

[2] See, e.g., Bornstein and Yaniv (1998) in the ultimatum game; Cox (2002) and Kugler et al. (2007) in the trust game; Bornstein et al. (2004) in the centipede game; Kocher and Sutter (2005) in a beauty-contest game; Luhan et al. (2009) in the dictator game. Cooper and Kagel (2005) find that groups play more strategically than individuals in signaling games. Another strand of literature in economics has studied the differences in equilibrium selection between groups and individuals in coordination games (Charness and Jackson, 2007; Croson and Fatas, 2009; Feri et al., 2010).

The fact that groups often perform better than individuals, in some quantifiable sense, suggests the possibility that consultation with a group, prior to an individual decision, might have analogous improving effects. However, such a conclusion may be premature. One reason for caution is the limited scope of direct evidence bearing on how deliberation within a group, prior to an individual decision, affects the quality of the latter. Two recent contributions which do directly address this, point to a positive impact of prior group deliberation on subsequent individual decisions. Maciejovsky et al. (2010) report that subjects who have solved decision problems as part of a group subsequently perform better as individuals in similar decision tasks. Charness et al. (2010) find that group consultation mitigates some decision anomalies found in individual choice experiments.[3] While these recent results chime with the broader literature comparing the success of individuals and groups, like many of the studies reviewed by Hastie (1986), both of them also share a design feature which may limit their generalizability: that feature is the use of tasks which have *demonstrably* correct solutions.

The use of tasks with correct solutions is a natural strategy to follow in any decision experiment where pursuit of the research question requires some unambiguous criterion of task success. But, the use of tasks with *demonstrably* correct solutions merits further consideration. We will say that the (correct) solution to a decision problem is *fully demonstrable*, in a given decision environment, when someone who knows the solution (and/or how to identify it) can convey that knowledge to any other individual facing the same decision.[4] In previous research, demonstrability has usually been implemented by using tasks which have correct answers that can be identified through the application of some commonly understood, or understandable, reasoning process (for instance, the task of finding the solution to a mathematical problem such as multiplying two numbers together; or that of identifying the solution to some logical reasoning problem, such as the famous 'selection task' of Wason, 1966). In such cases, while some individuals may not independently arrive at the correct solution, the task is demonstrable to the extent that they will *recognize* the solution when presented with suitable arguments to identify it.

[3] Specifically, they find significantly lower violations of the conjunction rule among decision-makers who are allowed to deliberate with other participants before submitting their answers.

[4] A richer discussion of the concept of demonstrability is provided by Laughlin and Ellis (1986). On their account, demonstrability requires four conditions: i) participants must reach a consensus on a common verbal or mathematical system; ii) there must be sufficient information within the system for reaching a solution; iii) the group members who are unable to reach a solution by themselves must have sufficient knowledge of the system to recognize and accept the correct solution; and iv) those who are able to reach the correct solution must have sufficient ability, motivation and time to demonstrate the solution to others.

High demonstrability of solutions may be an important ingredient explaining the relative success of groups over individuals across a range of existing experimental findings.[5] To appreciate this, imagine a decision environment with a task that has a correct solution and think of an individual decision maker as a degenerate group of size one. When the probability that at least one person in any group will know the solution increases with group size, then groups will inevitably have a higher probability of identifying the correct solution so long as solutions are demonstrable and those 'in the know' are willing to pass on what they know.[6]

But it is far from obvious that demonstrability is characteristic of most, or even many, of the settings in the world where individuals typically seek advice from others. Consider for instance, our opening examples related to financial, hiring or medical decisions. These types of decisions are qualitatively different from 'eureka-type' decision problems where a correct solution can be recognized via the application of a suitable system of reasoning. On the contrary, even 'experts' or 'professionals' often encounter difficulties in providing compelling arguments in defense of their estimations of, say, the profitability of a particular investment, the research publication potential of a newly appointed professor, or the risks associated with a new drug treatment. Indeed, such cases are often characterized by disagreements in the assessments of professionals and it can be difficult for someone seeking advice to decide who has the most accurate assessment, or perhaps equivalently, who is the most expert or informed advisor.[7] In part this may be due to the fact that the best alternative, ex post, may depend on some state of nature which is unknown at the point of decision. In fact, one might question whether, ex ante, it is meaningful to speak of there being a right answer at all in such cases. But even in many decision problems where it is meaningful to speak of there being a correct answer (e.g., a jury member trying to decide whether evidence against a defendant is compelling; or an analyst estimating the actual revenue of a company), there may not be any formal systems of reasoning

[5] This point is also made by, e.g., Hastie (1986), who reviews the social psychology literature comparing group and individual behavior in relation to tasks with varying degrees of solution demonstrability. Whereas groups clearly outperform individuals in tasks where solutions are easily demonstrated, the comparison is less clear in tasks where solutions are more difficult to demonstrate: here groups seem to perform slightly better than the average individual, but usually below the level of the best performing individual.

[6] In the context of environments that feature demonstrably correct solutions, the propensity for groups to 'outperform' individuals may thus be interpreted as at least a partly statistical phenomenon. It is a matter of ongoing debate, originating with Lorge and Solomon (1955), as to whether there is any improving effect of group membership beyond this (for more recent commentary on this point see Charness et al., 2010).

[7] As a practical illustration consider the difficulties interpreting the often conflicting advice about the risks of drinking alcohol during pregnancy (see, e.g., http://www.nhs.uk/news/2007/October/Pages/Pregnancyandalcohol.aspx).

that clearly identify, or unambiguously validate it. This suggests practical obstacles to demonstrability in potentially many cases of applied interest.

What effects could consultation have in environments where demonstrability is low? We propose a simple model of consultation which suggests that, whereas agents who consult will improve their chances of success whenever demonstrability is high, consultation within a group could *worsen* subsequent individual decision making when demonstrability is low. Consider a task with n possible answers and a uniquely correct solution.[8] There are three types of agents whom we label: 'expert', 'naïve' and 'mimics'. We consider two variants of the model, one with and one without consultation. In the model without consultation, each individual selects an action according to a fixed probability: 'experts' choose the correct answer with $q_{expert} = 1$; 'naive' agents choose at random selecting the correct solution with probability $q_{naïve} = 1/n$; and 'mimics' choose the correct solution with a probability intermediate[9] between that of the experts and the naive *(i.e.* $1/n < q_{mimics} < 1$*).* We model consultation as a two-stage decision process: in the first stage, each agent announces an intention generated according to the probability governing their own type; in the second stage, the experts and the naïve agents follow their intentions, but the mimics copy the actions of some subset of the other agents. We assume that the class of agents they copy depends on the demonstrability of the correct solution. Specifically, in decision problems with a fully demonstrable solution, we assume that mimics copy the 'experts' (on the grounds that experts can demonstrate the right answer) and hence select the correct solution with probability one. In decision problems without an obviously demonstrable solution, we assume that mimics will simply follow the majority intention.[10] In this case, it is easy to see that consultation can improve performance in populations where the 'experts' are sufficiently abundant, but in a population where the non-experts are prevalent, overall performance may worsen.

[8] For the purpose of maintaining an unambiguous benchmark to evaluate performance, in this paper we focus on decision environments that do have correct solutions.

[9] As motivation for this, think of choice among complex set of mobile phone tariffs where for any individual there is some best tariff, but it is difficult to identify. Experts know the best option for any individual, naïve agents choose without discriminating, whereas mimics represent partially informed individuals who, say, research options using a subset of available information.

[10] Laughlin and Ellis (1986) propose that the number of individuals necessary to reach a collective group decision is inversely related to demonstrability. A single correct group member is necessary and sufficient to reach a correct group decision in tasks with highly demonstrable correct answers (e.g. mathematical problems), whereas pluralities or majorities are needed in tasks with correct answers that are not obviously demonstrable (e.g. estimation tasks), or in 'judgmental' tasks that do not have a correct answer (e.g. attitudinal or aesthetic judgments).

This model illustrates the interesting possibility that consultation within a group could worsen individual decision making in tasks with uniquely correct answers, but ones that are low on demonstrability. Of course, our model is very simple and based on arguably naïve assumptions (e.g. about the following behavior of mimics). So an obvious question is whether such negative effects of consultation actually operate in real decisions. This is precisely the question addressed in the study we report here.

Our study focuses on an individual decision task that is designed to have a correct solution which is not obviously demonstrable. We observed individual decisions in this task comparing behavior across two treatments. In one treatment, before facing the decision, subjects had the opportunity to discuss the task with other participants. We compare decisions made by these subjects with participants in a control group who had no opportunity to discuss the task with others. Because, in both treatments we observe *individual* answers to the task, our study differs from of the literature which compares the decisions of groups with those of single individuals. Instead, our study focuses on whether deliberating with others has an impact on subsequent individual decision making.

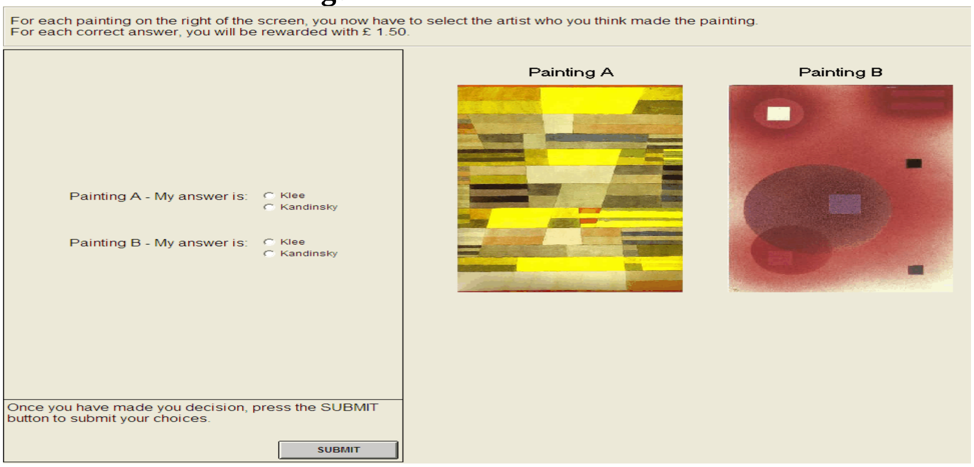
Section 2 describes the experimental design. We report our results in Section 3. Our primary findings are that subjects who discuss the decision task with their teammates believe that the deliberation has helped them, but they actually perform significantly worse and earn significantly less. The negative effects of group interaction on performance are particularly marked for females. Section 4 concludes.

2. EXPERIMENTAL DESIGN & PROCEDURES

Our decision task is designed such that i) there exists a correct solution known to the experimenter such that it is possible to establish a meaningful benchmark for performance; ii) at least some of the participants may be expected to have knowledge of the correct solution, such that there is scope to analyze the extent to which knowledge is transmitted via consultation; iii) crucially, solution demonstrability is low in the sense that proposed solutions cannot be validated via a formal system of reasoning that is commonly shared by participants. All of the subjects in our experiment faced a pair of decision problems in which they were asked to consider two paintings. For each painting subjects had to select which of two artists, Paul Klee or Wassily

Kandinsky, had made the painting.[11] Subjects had a clear incentive to answer correctly, if they thought they knew the right answer, because they received £1.50 for each correct answer (and nothing for any incorrect answer). Figure 1 shows the computer screen that subjects used to submit their answers.

Figure 1 – The Decision Task



We contend that this task has relatively low demonstrability[12] because, in comparison to tasks which have demonstrably correct answers, the solutions to our painting tasks cannot be identified via the application of any system of reasoning that would be commonly understood by our subjects.[13]

Each subject took part in one of two treatments which we refer to as the **Individual** and **Team** treatments. In the Team treatment, the decision task was preceded by a 'team discussion' stage in which subjects were randomly divided into three teams of six. After being assigned to a team, subjects had the opportunity to discuss the task with their teammates before submitting their answers. Team interaction lasted for 5 minutes during which time subjects could communicate with the rest of their team via a team-chat program. Subjects were told that they could use the

[11] The two paintings were *Monument in Fertile Country*, 1929 by Paul Klee (painting A in the experiment), and *Weighing*, 1928 by Wassily Kandinsky (painting B). A similar task is also used by Chen and Li (2009) as part of a social identity manipulation.
[12] Our task is similar in spirit to some of the low-demonstrability tasks used in the social psychology literature, such as world knowledge questions (e.g. what is the capital city of Lithuania) or estimation tasks (e.g. what is the population of China).
[13] We do not rule out the possibility that there may be arguments that a community of art experts might recognize as identifying the correct responses to our task (perhaps involving, for example, references to the style of brush strokes used by the different artists). Our subjects, however, were not selected on the basis of, or expected to have, any particular expertise in art history. Section 3 presents some analysis of how much knowledge our subjects in fact had, which supports this expectation.

chat-program to discuss the decision task and to get help from, or offer help to, other members of their team. They knew that messages were only shared among the members of their own team. At the end of the 5 minutes, subjects were shown the computer screen reproduced in Figure 1, and they individually submitted their answers. In the Individual treatment there was no 'team discussion' stage, and subjects were directly shown the decision task.

Note that in both treatments subjects made private decisions as individuals. Thus, our study provides a controlled test of whether being able to interact and discuss the decision task with others has an impact on subsequent individual performance as compared to a baseline situation (the Individual treatment) where team deliberation is not possible.

The decision-task reported here was the first part of a larger experiment and was followed by a one-shot sequential principal-agent game which is not related to the question examined in this paper. Subjects were informed at the beginning of the session that the experiment consisted of two parts, but detailed information about the second part was only given once everyone had completed the first part. At the end of the second part of the experiment, subjects completed a short post-experimental questionnaire eliciting basic demographic and attitudinal information. This included a self-assessment of subjects' risk and trust attitudes. Risk attitudes were elicited using the SOEP general risk question discussed in Dohmen et al. (2011). The question reads: "*Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?*", and subjects answered on a scale from 0 (risk averse) to 10 (fully prepared to take risk). Trust attitudes were elicited using the WVS Trust question ("*Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?*"), to which they replied either by saying that they believe that "*most people can be trusted*" or that one needs "*to be very careful in dealing with people*".[14] In the data analysis, responses on these two questions will enter as controls in a regression of subjects' responses to the painting task.[15]

At the start of the experiment subjects were randomly seated at visually separated computer terminals and were given a written set of instructions that the experimenter also read aloud (instructions are reproduced in Appendix A). The experiment was conducted using the software

[14] The average response to the SOEP risk question was 5.89 (s.d. 2.18). Responses to the WVS Trust question reveal that about 48% of our subjects believe that "*most people can be trusted*".

[15] While we do not have any especially strong priors here, the WVS trust measure might proxy the extent to which individuals are prepared to follow the advice of others and, on some conjectures, the SOEP risk measure might correlate with patterns in responses to the two tasks (for instance, a risk averse agent who believed that the two paintings were by different artists, but could not identify which painted which, may prefer to give the same answer in both tasks). In the analysis, however, we find no evidence that these measures have any predictive power.

z-Tree (Fischbacher, 2007). Subjects were students from a wide range of disciplines recruited through the online recruitment system ORSEE (Greiner, 2004). We conducted 19 sessions in total (15 in the Team treatment and 4 in the Individual treatment), with 18 subjects in each session. Consequently, we observed individual decisions of three-hundred and forty-two subjects: two-hundred-and-seventy subjects in the Team treatment and seventy-two in the Individual treatment. The average age of the participants was 20.2 years and 50% were female. No subject took part in more than one session. Subjects' earnings from the first part of the experiment ranged from £0.00 to £3.00, averaging £1.34. Subjects were paid in private and in cash at the end of each session.

3. RESULTS

3.1 Does consultation improve decision making?

Figure 2 shows the distribution of correct answers submitted in the Individual and Team treatments. In the Individual treatment (top panel of Figure 2) 38% of the subjects answer correctly to both painting questions, 33% answer correctly to one question and 29% submit two wrong answers. We note that if individuals chose entirely at random half of all subjects would be expected to get one question right, 25% of subjects would get both questions right and the remaining 25% would get none right. We can confidently reject the null that subjects in individual treatment choose at random (χ^2 goodness of fit test: $p = 0.000$). It is apparent from Figure 2 that fewer than the expected 50% of subjects get only one question right ($p = 0.006$ on a binomial test) and more than the expected 25% of subjects get two questions right ($p = 0.020$ on a binomial test). This shows that there is *some* knowledge of the correct answer in our subject pool, but knowledge is clearly far from perfect. Overall, these features of the distribution suggest that our task is satisfactorily calibrated for the purpose of investigating the extent to which imperfect knowledge is conveyed through group discussion.[16]

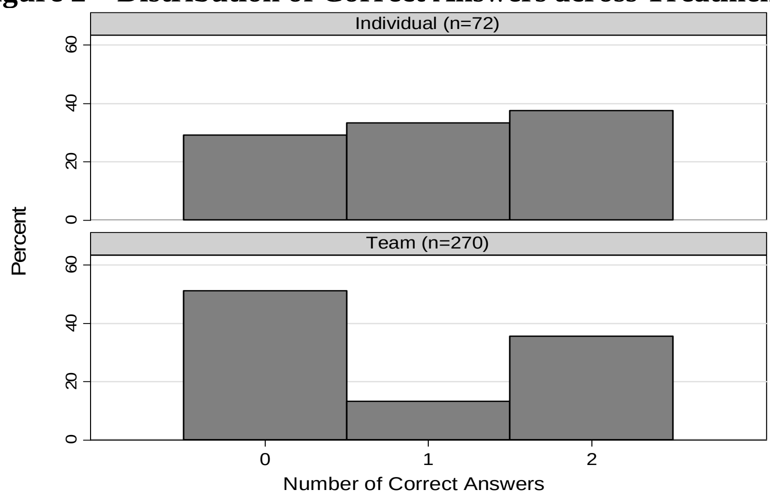
Turning to the Team treatment, note that, whereas the fraction of subjects answering correctly to both painting questions is similar to that in the Individual treatment[17] (36% in Team

[16] If choices in this treatment were either indistinguishable from random behavior, or very close to perfect, that could reduce the scope for observing any effects of group discussion that might be observable in less extreme environments.

[17] Note that, as in the Individual treatment, observations in the Team treatment are also decisions of single individuals, albeit preceded by team discussion.

vs. 38% in Individual), the two treatments differ markedly in the proportions of subjects with either zero or one correct answer. In Team, the proportion of subjects submitting two wrong answers is 51% (29% in Individual), while only 13% of subjects submit one correct answer (33% in Individual). We can strongly reject the hypothesis that the distribution of correct answers is the same across the two treatments: $\chi^2(2df) = 18.91$; $p < 0.001$. In the aggregate, subjects who participated in the group discussion were *less* successful (in terms of the average number of correct choices or, equivalently, the average earnings) than those who did not. In terms of earnings, subjects in the Team treatment earned, on average, 22% less than those in the Individual treatment, and the difference is significant at the 5% level according to a two-sided Wilcoxon rank-sum test (z = 2.14; p = 0.032).

Figure 2 – Distribution of Correct Answers across Treatments



We further examine the distributions of correct answers across treatments using regression analysis that allows us to control for observable differences across subjects. We use a generalized ordered logit regression model where the dependent variable describes whether a subject answers correctly to zero, one or two questions.[18] In Model I we only use a dummy variable assuming value 1 if the subject was in the Team treatment, and value 0 if the subject

[18] The generalized ordered logit regression model allows to relax the 'parallel regression assumption' of the standard ordered logit model whereby the coefficients that describe the relationship between, e.g., submitting zero correct answers versus submitting one or more correct answers are the same as those that describe the relationship between submitting two correct answers versus submitting zero or one correct answers, etc. (see, e.g. Long and Freese, 2006). This assumption is violated for the three regression models presented in Table 1 according to a Brant test (in all models p < 0.05). The test also showed that the largest violations were for the treatment dummy. Thus, we used the command gologit2 in STATA 11 to estimate generalized ordered logit regressions where the parallel regression assumption was relaxed for the treatment dummy and maintained for other regressors.

was in the Individual treatment. Model II expands Model I by adding controls for subjects' personal characteristics (gender, a dummy variable indicating whether a subject studies Humanities, and a self-assessment of the subject's risk and trust attitudes), and for session effects. Model III expands Model II by introducing interaction terms between the treatment dummy and the other regressors. Table 1 reports the regression results, displayed as factor changes in the odds of answering correctly.[19]

Table 1 – Regression Analysis of the Number of Correct Answers across Treatments

	I		II		III	
	At least one correct	*Both correct*	*At least one correct*	*Both correct*	*At least one correct*	*Both correct*
Team treatment	.394*** (.007)	.919 (.810)	.400** (.022)	.938 (.873)	.088*** (.007)	.228* (.086)
Male	-		.862 (.476)		.453* (.077)	
1 if Studies Humanities	-		.836 (.538)		.853 (.791)	
High Trust	-		.863 (.444)		.808 (.642)	
Risk Seeking	-		1.057 (.182)		.917 (.401)	
Session	-		.999 (.983)		.742 (.103)	
Team * Male	-		-		2.312* (.096)	
N.	*342*		*342*		*342*	
Wald χ2	*11.47*		*15.12*		*20.84*	
Prob > χ2	*0.003*		*0.034*		*0.053*	
Pseudo R^2	*0.025*		*0.029*		*0.039*	

Generalized ordered logit regressions. The dependent variable is the number of correct answers submitted by a subject. The results are displayed as factor changes in the odds of answering correctly. P-values are reported in parentheses. Standard errors are adjusted for intragroup correlation (a subject's team is used as the independent clustering unit; in the Individual treatment clusters have size 1 and coincide with the subject). A constant is included in all models, but omitted from the Table output. For the treatment dummy 'Team' the parallel regression assumption is relaxed and the models report both the factor change in the odds of answering correctly to at least one question, and the factor change in the odds of answering correctly to both questions. In Model II and III the reference subject type is: in the Individual treatment, female, studying a discipline other than Humanities, classified as someone who believes that one needs "to be very careful in dealing with people". Model III includes all possible interactions between the Team variable and the other regressors, but only interactions significant at the 10% level are reported in the Table. In Model III the 'continuous' variables (Risk Seeking and Session) are centered at their mean before being interacted with the treatment dummy. Significance levels: * 10% ; ** 5%; *** 1%.

[19] Note that a factor change greater than 1 implies a positive effect on the odds of answering correctly whereas a factor change smaller than 1 implies a negative effect.

Model I shows that being in the Team treatment reduces substantially (by a factor of 0.39) the odds of submitting at least one correct answer. The estimates reveal that, whereas in the Individual treatment we expect to find approximately 2.43 subjects submitting at least one correct answer for every subject who submits no correct answer, in Team the same statistic falls to only 0.96. This effect is significant at the 1% level. Being in the Team, however, has only a relatively small and insignificant impact on the odds of answering correctly to both questions (the odds are reduced only by a factor of 0.92, $p = 0.810$). This is consistent with the intuitively plausible idea that those who actually know more are less likely to be swayed by the crowd.

These results are robust to the inclusion of controls for personal characteristics and session effects (Model II). In particular, note how the negative effect of the treatment on the odds of submitting at least one correct answer remains large and statistically significant at the 5% level. The regression also shows that there is no clear relationship between any of the controls introduced in Model II and the propensity to answer correctly.

In Model III we interact the treatment dummy with the regressors used in Model II. Among the interaction terms, only the interaction between the treatment and the gender dummy is significant at the 10% level. Interestingly, the model reveals that being in the Team treatment is especially detrimental for female subjects: the odds of submitting at least one correct answer are reduced dramatically for female participants (by a factor of 0.09), and the effect is highly significant ($p = 0.007$). The effect is smaller for male participants (the odds decrease approximately by a factor of 0.20), and it is only marginally significant ($\chi^2(1df) = 3.45$; $p = 0.063$). Moreover, Model III shows that, for female subjects, being in the Team treatment also reduces the odds of answering correctly to both questions although the effect is only significant at the 10% level (the effect is insignificant for males).

3.2 The unrecognized curse of consensus

Why would the opportunity to consult with others have generated lower performance? In the introduction, as part of the motivation to our study, we sketched a simple model which generated that outcome. Central to our model was a tendency for individuals who know a little to be lured by a crowd that knows less.

A very striking feature of our data is a tendency for subjects in teams to give the same answers to the painting questions as those given by their teammates. In approximately 84% of teams, an absolute majority of team members submitted identical answers to the two questions.

In about a third of the teams (14 out of 45), all six team members submitted the same answers; in the other 24 teams a majority of four or five members submitted identical answers, with one or two 'outsiders' submitting different answers. This tendency arose even though participants submitted their answers *individually,* in private, and with no mention anywhere in the experimental instructions that a team had to reach a consensus.

A second interesting feature of team responses is that whether or not a team had reached a consensus is strongly associated with subjects' evaluations of whether communicating with the other team members was a helpful input to the decision task. At the end of the experiment, but before being informed about the outcome of the decision task, subjects in the Team treatment were asked to rate how much they thought that communicating with their team members had helped them solve the two painting questions. They responded on a scale from 1 ('not at all helpful') to 10 ('extremely helpful')[20] and from these responses we construct, for each team, a helpfulness index which is the mean of reported values for each team. For the 14 teams where every team member submitted the same answer to the painting questions, the average helpfulness index is 6.56. This falls to 4.75 in the 24 teams where there is a majority of four or five team members submitting identical answers, and it falls to 2.40 in the 7 teams where there was no solution submitted by an absolute majority. Both reductions are highly significant based on two sided-Wilcoxon rank-sum tests which produce $p = 0.003$ for the first comparison and $p = 0.004$ for the second.

The sting in the tail is that the poor performance of subjects in the Team treatment relative to those in the Individual treatment seems to be driven by those subjects who gave answers that are also submitted by the absolute majority in their team. If we exclude from the Team data all those subjects who formed part of a majority, we then find no significant difference between the average earnings of this subset (£1.438) and the average earnings of those in the Individual treatment (£1.625) (two-sided Wilcoxon rank-sum test: $z = 0.92$; $p = 0.355$). By contrast, average earnings in the Team treatment of those in majorities (£1.203) is significantly lower than the average earnings for the Individual treatment (two-sided Wilcoxon rank-sum test: $z = 2.43$; $p = 0.015$). Of course, it is possible that this inferior performance of majorities relative to subjects in the Individual treatment just reflects some form of sorting according to knowledge, which results in relatively low representation of the better informed individuals amongst majorities. If so, it

[20] The average response was 4.95 (s.d. 3.17).

would not necessarily suggest any negative performance effect of being associated with a consensus per se. But while such sorting might account for some of the differential for comparisons between subsets of subjects in the Team treatment and subjects in the Individual treatment, it cannot explain the *overall* differences in performance between treatments reported in the previous sub-section.

The face value interpretation of these data seems to be that consensus makes you feel good and perform worse. There is of course considerable evidence from social psychology that individuals have a strong tendency to form consensus, even when there is no basis for it.[21] But an intriguing question is why consensus answers perform poorly in our experiment; or, to put it differently, why majorities should have a tendency to coalesce around the wrong answers. While addressing this question takes us beyond the scope of the present study, we suggest a tentative hypothesis. It seems plausible to suppose that there may be a positive correlation between an individual's confidence that they know the answer and their willingness and/or ability to promote their guess to others. Coupling this with evidence showing a positive correlation between overconfidence and incompetence (e.g., Kruger and Dunning, 1999) would then suggest a possible mechanism at work in our data: teams tend to follow the advice of those who are both most confident and less informed.

4. CONCLUSION

We have reported an experiment designed to test the influence of consultation on individual decision making. Our work is partly motivated by an extensive background literature, briefly reviewed in the paper, which finds that groups often outperform individuals and we interpret our study as probing the *conditions* under which group interaction improves decision making. As we noted, most of the evidence supporting conclusions of the form 'teams make you smarter' or 'two heads are better than one' comes from experiments comparing decisions made by groups versus decisions made by individuals. Such comparisons are intrinsically interesting, not least because many important decisions *are* made by groups (e.g. decisions of the UK Monetary Policy Committee; budgetary allocation decisions by the World Bank's Board of Executive Directors; etc.). However, we have argued that many interesting and important decisions where groups may play a role are better construed as *individual decisions* that involve an element of

[21] For discussion of some early evidence see Sherif (1935).

consultation and part of our objective was to examine the extent to which the beneficial effects established for decisions made by groups extend to decisions made by individuals who consult.

A second distinguishing feature of our experiment was the use of tasks with (correct) solutions that are low on *demonstrability*. This design feature had a number of connected motivations discussed in the introduction and reviewed here. The first stems from recognizing that the bulk of evidence pointing to beneficial effects of group decisions might be partly a by-product of experimental designs featuring tasks with high demonstrability. Indeed, when tasks have fully demonstrable solutions then, by definition, those who have knowledge of them can convey their knowledge to others. As such, adopting tasks with low demonstrability can be seen as providing a tougher test of the extent to which the knowledge possessed by some members of a group can be successfully transmitted to other members of it. That test is relevant, not least, because many interesting decisions in the world – and, in particular, many of those where consulting is commonplace - tend to have low demonstrability. A final motivation flows from our simple model which illustrates how consultation could have perverse effects when demonstrability is low.

Our primary finding is that beneficial effects of group participation do not extend to our environment. Our benchmark (Individual) treatment established that our subject pool had some knowledgeable participants. Yet those who had the opportunity to consult, on average, performed worse and earned less than those who did not. This negative effect of consultation seems to reflect a tendency for individuals to follow the, relatively poorly informed, crowd and it was particularly marked for females.

REFERENCES

Bone, J., J. Hey, and J. Suckling. 1999. Are Groups More (or Less) Consistent Than Individuals? *Journal of Risk and Uncertainty* 18(1), 63-81-81.

Bornstein, G., and I. Yaniv. 1998. Individual and group behavior in the ultimatum game: Are groups more "rational" players? *Experimental Economics* 1(1), 101-108.

Bornstein, G., T. Kugler, and A. Ziegelmeyer. 2004. Individual and group decisions in the centipede game: Are groups more "rational" players? *Journal of Experimental Social Psychology* 40(5), 599-605.

Charness, G., and M.O. Jackson. 2007. Group play in games and the role of consent in network formation. *Journal of Economic Theory* 136(1), 417-445.

Charness, G., E. Karni, and D. Levin. 2007. Individual and group decision making under risk: An experimental study of Bayesian updating and violations of first-order stochastic dominance. *Journal of Risk and Uncertainty* 35(2), 129-148.

Charness, G., E. Karni, and D. Levin. 2010. On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior* 68(2), 551-556.

Chen, Y., and S.X. Li. 2009. Group Identity and Social Preferences. *American Economic Review* 99(1), 431-457.

Cooper, D.J., and J.H. Kagel. 2005. Are Two Heads Better Than One? Team versus Individual Play in Signaling Games. *American Economic Review* 95(3), 477-509.

Cox, J. 2002. Trust, Reciprocity, and Other-Regarding Preferences: Groups vs. Individuals and Males vs. Females. In *Advances in Experimental Business Research*, ed. Rami Zwick and Amnon Rapoport. Kluwer Academic Publishers.

Croson, R.T.A., and E. Fatás. 2009. The Impact of Consensus: A Comment on "Efficiency Gains from Team-Based Coordination—Large Scale Experimental Evidence". CBEES Working Paper 09-12.

Dohmen, T.J., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G.G. Wagner. 2011. Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences. *Journal of the European Economic Association* 9(3), 522–550.

Feri, F., B. Irlenbusch, and M. Sutter. 2010. Efficiency Gains from Team-Based Coordination—Large-Scale Experimental Evidence. *American Economic Review* 100(4), 1892-1912.

Fischbacher, U. 2007. z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics* 10(2), 171-178.

Greiner, B. 2004. An Online Recruitment System for Economic Experiments. In *Forschung und wissenschaftliches Rechnen GWDG Bericht 63*, 79-93. Göttingen: Gesellschaft für Wissenschaftliche Datenverarbeitung.

Hastie, R. 1986. Review essay: Experimental evidence on group accuracy. In *Information pooling and group accuracy*, ed. G. Owen and B. Grofman, 129-157. Westport, CT: JAI Press.

Kocher, M.G., and M. Sutter. 2005. The Decision Maker Matters: Individual Versus Group Behaviour in Experimental Beauty-Contest Games*. *The Economic Journal* 115(500), 200-223.

Kruger, J., and D. Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology* 77(6), 1121-1134.

Kugler, T., G. Bornstein, M.G. Kocher, and M. Sutter. 2007. Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy. *Journal of Economic Psychology* 28(6), 646-657.

Laughlin, P.R., and A.L. Ellis. 1986. Demonstrability and social combination processes on mathematical intellective tasks. *Journal of Experimental Social Psychology* 22(3), 177-189.

Laughlin, P.R., E.C. Hatch, J.S. Silver, and L. Boh. 2006. Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology* 90(4), 644-651.

Laughlin, P.R., M.L. Zander, E.M. Knievel, and T.K. Tan. 2003. Groups Perform Better Than the Best Individuals on Letters-to-Numbers Problems: Informative Equations and Effective Strategies. *Journal of Personality and Social Psychology* 85(4) (October), 684-694.

Long, S.J., and J. Freese. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. Second Edition. College Station, Texas: Stata Press.

Lorge, I., and H. Solomon. 1955. Two models of group behavior in the solution of eureka-type problems. *Psychometrika* 20(2), 139-148.

Luhan, W., M.G. Kocher, and M. Sutter. 2009. Group polarization in the team dictator game reconsidered. *Experimental Economics* 12(1), 26-41.

Maciejovsky, B., M. Sutter, D.V. Budescu, and P. Bernau. 2010. Teams Make You Smarter: Learning and Knowledge Transfer in Auctions and Markets by Teams and Individuals. IZA Discussion Paper No. 5105.

Munro, A., and D. Popov. 2009. A Missing Link in Behavioural Economics? A Portmanteau Experiment on the Relevance of Individual Decision Anomalies for Households. Royal Holloway College Discussion Paper 09-10.

Rockenbach, B., A. Sadrieh, and B. Mathauschek. 2007. Teams take the better risks. *Journal of Economic Behavior & Organization* 63(3) (July), 412-422.

Sherif, M. 1935. A study of some social factors in perception. *Archives of Psychology* 187(27), 1-60.

Sutter, M. 2007. Are teams prone to myopic loss aversion? An experimental study on individual versus team investment behavior. *Economics Letters* 97(2) (November), 128-132.

Wason, P.C. 1966. Reasoning. In *New horizons in psychology*, ed. B.M. Foss. Penguin. Harmondsworth.

APPENDIX A (not for publication)

Preliminary Instructions [common in both treatments]

Welcome! You are about to take part in a decision-making experiment. This experiment is run by the "Centre for Decision Research and Experimental Economics" and has been financed by various research foundations.

There are other people in this room, who are also participating in this experiment. All participants are reading the same instructions and have been recruited in the same way. Likewise, everyone is participating in this experiment for the first time. It is important that you do not talk to any of the other participants during the experiment. If you have a question at any time, raise your hand and a monitor will come to your desk to answer it.

This experiment consists of two parts: PART 1 and PART 2.

In each part you will be asked to make one or more decisions. Decisions made in one part of the experiment will not affect decisions or earnings in the other part.

You will be informed of any outcome (including your earnings) from PART 1 and PART 2 of the experiment only at the end of the session. Therefore everyone will make their decisions in PART 2 of the experiment without knowing any outcome from PART 1.

Your earnings will be paid to you in private and in cash at the end of the experiment.

We will shortly give you instructions about PART 1 of the experiment. You will receive instructions about PART 2 once everyone in the room has completed PART 1 of the experiment.

Instructions for the Decision Task [Individual Treatment]

In PART 1 of the experiment you will be shown a screen with two paintings by two modern artists: Paul Klee and Wassily Kandisky. Your task in PART 1 of the experiment is, for each painting, to select the artist who you think made the painting. For each correct answer, you will be rewarded with £1.50. You will be paid in private and in cash at the end of the experiment.

Please raise your hand if you have any questions.

Instructions for the Decision Task [Team Treatment]

In PART 1 of the experiment you and 5 other participants will be randomly assigned to a group. In a moment, the experimenter will bring around a bag containing 18 plastic balls and ask each of you to randomly draw a ball from the bag. 6 balls in the bag have a SQUARE drawn on them, 6 have a TRIANGLE drawn on them, and 6 have a CIRCLE drawn on them. Depending on your draw you will be assigned either to the SQUARE group, to the TRIANGLE group or to the CIRCLE group. Likewise, all other participants in the room will be assigned to a group according to their draw. Thus, there will be 5 other participants in your group. You will not learn the identity of the other participants in your group, during or after today's session.

Once every participant has been assigned to a group, a screen will appear showing you two paintings by two modern artists: Paul Klee and Wassily Kandisky. Your task in PART 1 of the experiment is, for each painting, to select the artist who you think made the painting. For each correct answer, you will be rewarded with £1.50. You will be paid in private and in cash at the end of the experiment.

Before you begin making your choices, you will have 5 minutes to use a group chat program to get help from or offer help to other members in your own group. Messages will be shared only among the members of your own group. You will not be able to see the messages exchanged among the members of the other groups. People in the other groups will not see the messages from your group. Except for the following restrictions, you can type whatever you want in the lower box of the chat program.

Restrictions on messages

1. You must not identify yourself or send any information that could be used to identify you (for example, your name, contact details or seat in the room);
2. You must not make any threats, insults or use any obscene or offensive language.

If you violate these rules your payment will be forfeited.

Please raise your hand if you have any questions.