

Persuasion: An experimental study of team decision making

November 17, 2015

This paper studies persuasion within teams and investigates why teams commonly take, by some measures, better decisions than individuals. The analysis is based on data from electronic communication within teams of two players. Thanks to the experimental design, changes of an individual's decision can be attributed to the content of the team partner's message. The results for knowledge-related and strategic problems show that individuals' decisions change upon receiving more informative and sophisticated arguments and remain the same otherwise. This individual behavior is an essential part of the information aggregation in teams and can explain the advantage of teams in decision making and in games.

JEL Classification: D70, C92, D83

Keywords: Persuasion, Team decision making, Information aggregation, Level- k model

1. Introduction

In this study, I investigate persuasion when the communicating agents have a common interest to take good decisions but have limited information or are boundedly rational. Studying this kind of persuasion describes the fundamental mechanisms within important economic institutions such as the decision making team or other networks of advice and consultation. Deliberation in committees, boards and groups as well as consultation is ubiquitous in business, politics and private life. A large and growing literature on team decision making in psychology, management and behavioral economics suggests that communication with others leads to, by some measures, better decisions. The purpose of this paper is to understand how teams achieve this. The focus will be on individual team members and the circumstances under which they get persuaded.

I define persuasion as a change in a receiver’s preferred action resulting from the exposure to a sender’s argument. By the nature of this phenomenon, a rigorous study requires detailed experimental data. The communication between individuals needs to be free and incentivized to be comparable to naturally occurring environments. At the same time, the content and originator of an argument needs to be tracked as well as its effect at the destination in terms of pre- and post-communication decisions. This paper makes use of an experimental design, introduced by Burchardi and Penczynski (2014), that provides such data by having two participants play as one team and communicate electronically in the following way: Both players individually suggest a decision and justify it in a written message. After the simultaneous exchange of suggestion and message, the team partners finally decide individually on the action they want the team to take. One of the two final decisions is randomly implemented as the team’s action.¹ Hence, with a probability of 0.5 the partner and not the player herself determines the team’s action, a fact which provides incentives to argue convincingly in the message.

With this experimental design, the paper is – to the best of my knowledge – the first to use such a clean message structure to investigate persuasion and team interaction. It does so in two important contexts. First, teams are confronted with a purely knowledge-related problem as they have to answer trivia questions about the year of a historic event. The second setting involves strategic interaction in the “beauty contest” and the “hide and seek” games. In the first game each team chooses a number between 0 and 100. The team closest to $2/3$ of the average of all teams’ numbers wins (Nagel, 1995). The fine action space and the game’s dominance solvability make it very useful for investigating persuasive strategic arguments. In the constant-sum hide and seek game a hider team hides a treasure and a seeker team seeks the treasure at one of four locations. The team that holds the treasure at the end wins a prize (Rubinstein and Tversky, 1993). Despite a coarser action space, this second game allows to test the main result in a normal-form game with unique mixed strategy equilibrium.

¹With the implementation of a random dictator, the study abstracts from the role of voting rules and decision aggregation in teams, as studied, for example, in Charness and Jackson (2007).

For the use in my analysis, the messages are classified. In the trivia questions, the information content of each message is quantified. In the strategic settings, the messages are classified along the lines of the level- k model of reasoning, a procedural model of strategic thinking (see Nagel (1995); Stahl and Wilson (1995); Costa-Gomes and Crawford (2006)). The level of reasoning is defined as the number of iterated best responses applied to a belief about non-strategic play.

I analyze the effect of the received message on the final decision. For the trivia questions, I find that messages with a relatively higher number of informative facts or a more narrow range of possible years are more persuasive. This explains the more concentrated distribution of final decisions and provides a clear understanding of the process of information aggregation in teams.

In the strategic settings, relatively more sophisticated messages are more persuasive. In other words, arguments of a higher level of reasoning induce more changes in the decisions of players that reason on a lower level than lower level arguments in the decisions of higher level players. This effect is robust across both games. It provides an explanation as to why groups' strategic decisions are often found to be more rational than individuals' decisions (see surveys by Kugler, Kausel and Kocher, 2012; Charness and Sutter, 2012). In a population that is heterogeneous in the levels of reasoning, this effect makes higher level arguments more likely to shape the decision of a representative team.

Furthermore, I find less sophisticated arguments that endorse unexpected decisions to be persuasive in the sense that they induce more sophisticated players to change their decisions in the direction of the unexpected decision. This effect can be interpreted as an update of sophisticated players' beliefs about lower level decisions. In a symmetric game, less sophisticated team partners provide information about the likely reasoning and choices of opponents, so that sophisticated players come to hold beliefs about the population that are based on more information.

Overall, the results show that individuals give up their initial decision when confronted with more sophisticated arguments and maintain their initial decision otherwise. Such a selection process is an important mechanism to improve decisions when arguments are pooled and team members are heterogeneous with respect to individual knowledge or strategic sophistication. While team decision making processes are certainly influenced positively and negatively by many factors, my experimental study illuminates one mechanism that shows how teams aggregate information and strategic arguments and explains why teams take, by some measures, better decisions.

2. Literature

In economics, persuasion is commonly modelled as strategic transmission of information from a sender to a receiver with different interests. The sender reveals or withholds private information

in order to persuade the receiver into taking specific actions by changing his beliefs (see Crawford and Sobel, 1982; Milgrom and Roberts, 1986; Glazer and Rubinstein, 2004; Kamenica and Gentzkow, 2011). Many models assume rational Bayesian behavior but also various kinds of non-Bayesian behavior have been analyzed in this context (DeMarzo, Vayanos and Zwiebel, 2003; Mullainathan, Schwartzstein and Shleifer, 2008; Glazer and Rubinstein, 2012).

These models are not intended to make predictions about persuasion processes when parties have aligned interests. For that purpose, it is fruitful to use models of limited information and bounded rationality. To my knowledge, such kind of persuasion has not been formalized. Possible starting points are general models of limited knowledge and bounded rationality such as Modica and Rustichini (1994) and Lipman (1999). For the purposes of this study I derive predictions directly from the respective models of knowledge (see Rubinstein, 1998) and strategic thinking (Nagel, 1995; Stahl and Wilson, 1995).

Another source of predictions is the psychology literature on persuasion. The “persuasive argument theory” views group decisions as influenced by the number and persuasiveness of arguments considered (Thorndike, 1938; Burnstein, Vinokur and Trope, 1973). The team’s task or problem itself is found to have a significant influence on the team decision process. Whether the solution is or is not easily demonstrable divides tasks into “intellective” and “judgmental” ones (Laughlin and Ellis, 1986). In psychology, the common finding on various tasks is that groups outperform individuals on average, especially in intellective tasks (“eureka problems”, Hill, 1982; Hastie, 1983; Davis, 1992; Levine and Moreland, 2006).

A growing literature in experimental economics reports results of groups deciding more rationally than individuals (see Kugler, Kausel and Kocher, 2012; Charness and Sutter, 2012, for a detailed overview). As one of the first studies, Bornstein and Yaniv (1998) look at teams of three players in an ultimatum game and find that actions of both proposers and responders are closer to the subgame-perfect Nash equilibrium when teams rather than individuals take decisions. Blinder and Morgan (2005) run two experiments, one abstract urn composition inference experiment and a monetary policy experiment and find in both cases that groups make better decisions. In the beauty contest game, Kocher and Sutter (2005) find that team reasoning leads to a lower variance in the actions chosen and an accelerated convergence to the Nash equilibrium in repeated rounds. A substantial literature on advice is related to similar processes and finds that advisees take decisions closer to equilibrium (Schotter, 2003; Çelen et al., 2010).

The initial method to investigate team decision making has been the observation of team decisions without an evaluation of the team communication. The use of electronic messaging makes the content of team communication easily available for analysis, as pioneered by Cooper and Kagel (2005) who use data on chat dialogues to identify factors that promote strategic play in signaling experiments. This method has been and continues to be used for various games (Luhan, Kocher and Sutter, 2009; Casari, Zhang and Jackson, 2012; Sheremeta and

Zhang, 2010). A different approach that leads to similar kinds of data is to record or videotape spoken group discussions and transcribe the communication afterwards as pioneered in the experimental economics literature by Bosman, Hennig-Schmidt and Van Winden (2006). In all these cases, dialogues within teams reveal arguments that shape the group decision. They make it difficult, however, to identify and track the changes of individuals' positions. The communication protocol in this study features the elementary team discussion of two members simultaneously exchanging one message each. As this permits to cleanly observe both the source of an argument and the effect at its destination, it is well suited to investigate persuasion on the individual level.

Parts of the data used here have been analyzed in Burchardi and Penczynski (2014) and Penczynski (2014) with a focus on individual reasoning, illuminating the nature of level-0 actions and beliefs and estimating the level- k distribution. The present study shows that their experimental design with intra-team communication is equally well suited to investigate very different questions, namely the mechanisms behind persuasion under aligned interests.

3. The experiments

3.1. Procedures

The experiment on the trivia questions was conducted at the LEEDR Laboratory of Cornell University. In nine sessions, 138 students participated and answered three trivia questions about the year of a historic event, resulting in 414 observations of years between 0 and 2000. For each question, the students were randomly matched with a partner to form a team of two players. For each question, the team closest to the correct year received a prize of \$10 (\$5 per member). Each participant received a show-up fee of \$5.

The experiments on the beauty contest and the hide and seek games were conducted in the Experimental Economics Laboratory of the Department of Economics in Royal Holloway (University of London). In six sessions, the beauty contest game was played by 84 participants. In eight sessions, the hide and seek game was played by 114 participants. Each participant of the experiment was paid a show-up fee of £5 and the winning team won a prize of £20 (£10 per team player) in the beauty contest game and £10 (£5 per team player) for winning the hide and seek game. In all experiments, the participants were recruited by the host institution and consisted mainly of undergraduate students.

3.2. Design

The main feature of the experimental design is the communication protocol that structures the team partners' interaction. Individuals are randomly assigned into teams of two players. For

a given strategic situation, each player makes suggestions for the team action at two points in time. First, the so-called “suggested decision” and a justifying written message are exchanged between the team partners simultaneously.² After this, the “final decision” is taken individually. The computer chooses randomly one of the two final decisions to obtain the “team’s action.”³

The two individually made decisions – before and after the message exchange – indicate how the communication changes the subject’s preferred action. The message exchange itself presents a very simple version of group discussion in which arguments are pooled. The random choice to pin down the team’s action is a fixed decision rule which the teams do not need to agree on and which gives individual participants incentives to state their reasoning fully and clearly. The messages give an account of how individuals reason. It is thus known which arguments an individual puts forward as well as which she receives.

4. Persuasion among boundedly rational players

In order to fix ideas about persuasion in the context of bounded rationality and aligned incentives, I deduce hypotheses about characteristics of persuasive arguments from theoretical considerations.

Limited knowledge or bounded rationality is reflected by the endowment of the sender and the receiver with a set of considerations which represent how they think about the current problem. The sender can send a message to the receiver and thus transfer the specific considerations that are the basis for her proposed action. After the reception, the receiver can make use of novel considerations and possibly change her decision. I define a consideration to be persuasive if it induces such a change of the decision.

For the trivia questions, there are two ways to think about these considerations. The first way results from the idea that the information content and the number of considerations in the sender’s and receiver’s message are relevant for persuasion. Therefore, one could hypothesize that the higher is the number of transmitted informative considerations the higher is the probability that the receiver is confronted with persuasive arguments.

A second way to think about the epistemic questions follows from the set-theoretical model of knowledge commonly used in the study of bounded rationality (see Rubinstein, 1998, and references therein). In this model, “more information” is always reflected by ruling out states of the world. Since more information provides a better basis for good decisions, arguments that rule out more states of the world should be more persuasive. An empirical investigation of this model requires to quantify the range of years that the player still considers possible when presenting her arguments.

²Only one message exchange is possible. This keeps arguments attributable to the originating team member. However, no space or time restriction limited the number and scope of arguments.

³The experiment was programmed and conducted with the software z-Tree (Fischbacher, 2007).

While the epistemic questions relate to easily observable and quantifiable measures, the considerations that are expected in strategic thinking are less directly observed. In order to identify those, I turn to the level- k model of strategic reasoning and obtain categories of possible arguments (see Nagel, 1995; Stahl and Wilson, 1995; Camerer, Ho and Chong, 2004). This theory puts forward a specific reasoning process and thus is useful to identify different dimensions along which arguments are made. Ample empirical evidence shows that this way of modeling strategic reasoning describes the actual thinking processes very well (see Bosch-Domènech, Montalvo, Nagel and Satorra, 2002; Costa-Gomes and Crawford, 2006; Burchardi and Penczynski, 2014).

The model assumes that players apply a certain number of iterated best responses to a belief about others' non-strategic behavior. Player types are distinguished by the level of reasoning, a natural number which corresponds to the number of iterated best responses. Level-0 players do not reason strategically, they are modeled to play randomly and follow a probability distribution over the action space. Level-1 players apply one best response to their level-0 belief. Level-2 best responds to level-1 and so on. With this model in mind, three kinds of considerations can be distinguished.

First, the level of reasoning is the most obvious feature of an argument. Arguing about the level of reasoning is a more intellectual than judgmental process in the sense that players can potentially demonstrate that additional best responses help outperform those opponents who do not make those steps of reasoning. One can thus expect that higher sophistication in a subject's reasoning leads others to follow her actions.

Second, the level-0 belief describes how a player expects non-strategic play to be. Intuitively, it is difficult to demonstrate that a certain belief is more pertinent than another since this evaluation depends on each subject's experience and intuition. Therefore, this dimension of the argument is rather judgmental. Third, the level-0 action describes the decision a non-strategic player takes. This is comparable to the level-0 belief, the level-0 actions are then similarly judgmental. The working hypothesis will be that similar level-0 judgments are more persuasive than differing ones.⁴

5. Classification of communication transcripts

The communication messages are classified by the message characteristics that are relevant to the persuasiveness of a message as discussed in section 4. The following two sections describe how the research assistants proceeded.

⁴Unfortunately, due to the coarse and nominal action space in the hide and seek game, this hypothesis about the level-0 is only testable in the context of the beauty contest game.

5.1. Trivia questions

In order to quantify the informational content of the messages, two research assistants (RAs) separately read the messages and classified them according to the following criteria. The first three criteria follow from section 4, the fourth was included for completeness after reading messages from previous, unincentivized sessions. The criteria are i) the number of informative considerations, ii) the number of neutral (non-informative) considerations, iii) the precision of the guess in years, and iv) the reference to a credential by the participant.⁵ Since all classifications are simple quantifications, the averages are used in the analysis when the two RAs' individual classifications disagree. Table 1 gives summary statistics of the classification data and indicates that in all categories, classifications coincide in more than 90% of the messages.

The considerations were counted and distinguished into informative and neutral ones, depending on their relevance for determining the answer to the question. Practically, the RAs counted the number of facts a subjects mentions and distinguished facts that are useful or useless for determining the year in question. For example, regarding the year of birth of Thomas A. Edison a subject wrote: "I am not very sure of my suggested decision. I would guess the light bulb was introduced around 1830 and I would guess he was 30 or 40. I also believe he was born after the revolution." This is classified as containing three informative and no uninformative considerations. Table 1 indicates that on average a message contains 0.513 informative considerations and a very small number of non-informative considerations.

The precision of the guess results from ranges given by subjects for the year in question and indicates the number of years in the mentioned range. If a subject writes: "I would say that Elvis was born in the 20s or 30s", this indicator is 20 because the interval mentioned contains 20 years. For more implicit references to decade (10), early, late, or mid century (33), century (100), etc., guiding numbers were given to the RAs (in brackets). The default was the non-informative full range of 2000.

Finally, subjects might refer to a course they attended or a nationality that makes them expert in the question. Since such information might have an influence on the persuasiveness of the message, the RAs were asked to indicate these credentials as well. Table 1 shows that this is used very rarely. The table finally reports the length of the message which on average is 76 characters, about half a text message on a mobile phone.

⁵A fifth criterion is a player's self-qualification of his arguments, for example, whether he indicated to "have no idea" or "not to be sure", etc. Differences in this dimension are predictive of the persuasiveness of a message. It is, however, not further investigated since it follows – like the team partner's persuasion – from underlying information as captured by i) and iii), which are highly correlated with this fifth criterion.

Table 1: Summary statistics for message classification.

Variable	Mean	St. Dev.	Median	Min	Max	N	Agree (%)
Informative considerations	0.513	0.779	0	0	3.5	414	94.4
Neutral considerations	0.030	0.155	0	0	1	414	97.6
Range of guess	1350	913.7	2000	1	2000	414	92.5
Credential	0.012	0.109	0	0	1	414	98.3
Message characters	75.79	79.83	57.5	0	598	414	–

Notes: The last column gives the percentage of agreement between classifying RAs.

5.2. Strategic games

For both games, I draw on data from the classification process in Burchardi and Penczynski (2014) and Penczynski (2014). In the following, I describe the features of the classification relevant to the present study.

Two RAs separately read the messages and classified them according to two criteria, the level-0 beliefs or actions and the level of reasoning. The RAs were instructed to consider as level-0 type a player whose message does not exhibit “any strategic reasoning whatsoever.” This might consist of choosing a number randomly or based on non-strategic considerations such as taste. For this classification it is important that the player is not in any way best-responding to what he thinks others would do.

If the argument contains a belief about others’ play that serves as a starting point for best responses, but is in itself not derived by a best response or other strategic reasoning, the RAs note this as level-0 belief.

The level of reasoning corresponds to the applied number of iterated best responses to the level-0 belief. Since the messages might not precisely reveal each player’s level without remaining doubt, the RAs indicated a lower and an upper bound on the number of level of reasoning. These are defined as the lowest number of iterated best responses which are clearly stated and the highest number which can possibly be interpreted into the message.

For the purposes of this paper, I exclusively use the lower bounds from the classification. They reflect the strategic argument that was most easily accessible to subjects in the experiments because it was clearly contained in the message and most likely to be comprehended. Furthermore, I assign unclassified, empty messages a lower bound of 0 since their strategic content is naught.⁶

The classification instructions were self-contained and not complemented by verbal comments.⁷ In order to test the classification for stability and replicability, six additional RAs

⁶In the classification, there are 50 subjects with a difference of lower and upper bounds of 0, 19 with a difference of 1, and 1 with a difference of 2. 8 subjects are not given an upper bound, 6 are given no bounds due to empty messages (See Burchardi and Penczynski, 2014).

⁷After considering each part individually, the RAs met personally to reconcile their judgments and provided a joint classification.

individually classified the round 1 data. Table 2 shows that for 90% of the beauty contest messages and for 74% of the hide and seek messages, six or more out of eight RAs put the exact same level lower bound. The messages are an excellent coordination device and the classification procedure is shown to be very well replicable.

Table 2: Number of subjects for which the eight RAs’ level lower bounds classifications coincide maximally a given number of times. Empty messages are excluded.

		Coinciding RA level classifications							Total
		8	7	6	5	4	3	2	
BC	Nr. of subjects	40	15	15	3	5	–	–	78
	Cum. fraction	0.51	0.71	0.90	0.94	1.00	1.00	1.00	
H&S	Nr. of subjects	42	19	12	14	10	2	–	99
	Cum. fraction	0.42	0.62	0.74	0.88	0.98	1.00	1.00	

6. Results for the trivia questions

The teams faced three rounds of trivia questions about the years of historic events. The action space $A = \{0, 1, 2, \dots, 2000\}$ is very informative about revisions of earlier decisions. One question, for example, was: “Which is the year of birth of Nicolaus Copernikus, the first astronomer to formulate a scientific heliocentric cosmology?” A representative example of the ensuing communication follows.

Subject 14 (team 8, suggested decision 1456, final decision 1630):

*I have absolutely no idea, but I figured it might have happened in the early 1400’s.
It might have been during the renaissance era, but I don’t know when that is either.
I’m figuring the earlier the year the better, but I really couldn’t say.*

Subject 27 (team 8, suggested decision 1630, final decision 1630):

I think most of his work happened in the 1650s-1660s so I am guessing he was around 20 -30 something at that time.

This dialogue shows how the statement of the second team member induced the first one to adapt the decision. The correct answer is 1543.

6.1. The effects of communication

In a first step it has to be shown that the communication has an impact on the individual decisions and the message exchange approximates a group discussion. For that purpose, table 3 reports summary statistics for the distance between decisions and the correct answer. The final decision after the communication is on average about a year closer to the truth than the suggested decision before the communication; the median distance is 2.5 years smaller. Neither

of the two effects are significant.⁸ However, the standard deviation of the final decisions is significantly lower than the one of the suggested decision (F -test, $p < 0.001$) which shows that the group discussion disciplines outlier opinions.

Table 3: Distance to the truth before and after the team communication.

Distance to truth	Mean	St. Dev.	Median	Min	Max	N
Suggested decision	29.13	145.3	6	-973	526	414
Final decision	27.92	123.4	3.5	-973	505	414

The communication consists of two elements: (i) the player sending her suggested decision and (ii) justifying it by one written message. I first take a look at the influence of the partner’s proposal.

Disagreement is defined as the difference between the partner’s suggested decision and the player’s own suggested decision. It is positive if the partner’s suggestion is higher than the own one. Correction is defined as the difference between the own final decision and the own suggested decision. This variable is negative if one chooses a lower year after the communication. The signs of these variables are informative as they reflect the direction towards the partner’s decision and the direction of the individual’s revision.

Table 4: Directions of disagreement and correction in the trivia questions.

		<i>Disagreement</i>			Total
		neg.	0	pos.	
<i>Correction</i>	neg.	144	0	7	151
	0	52	12	46	110
	pos.	4	2	147	153
Total		200	14	200	414

Table 4 shows that a disagreement is mostly followed by a correction in the same direction. This way, there must be some information in the suggested decision or in the message that leads to a revision of the own decision by a large majority of the subjects. In order to predict which subjects correct and to which extent, we need to differentiate this analysis by the characteristics of the message. For that purpose, I quantify the effect of the communication on a decision more gradually. The so-called “relative correction” gives the correction divided by the disagreement. It takes the value 0 when no correction has been made and 1 when the partner’s proposal has been adopted entirely.⁹

The earlier considerations predict that there is a higher relative correction if the partner delivers more informative considerations. To test this hypothesis, the data is split up by the

⁸The small absolute values of the median reflect the “wisdom of crowds” as popularly covered by Surowiecki (2005).

⁹For example, imagine the suggested decision of player 1 being 1900 and the one of player 2 being 1800. Player 1 exhibits a disagreement of -100 . If she then chooses 1850 as a final decision, her correction is -50 and the relative correction is 0.5.

relative amount of informative considerations in the message.

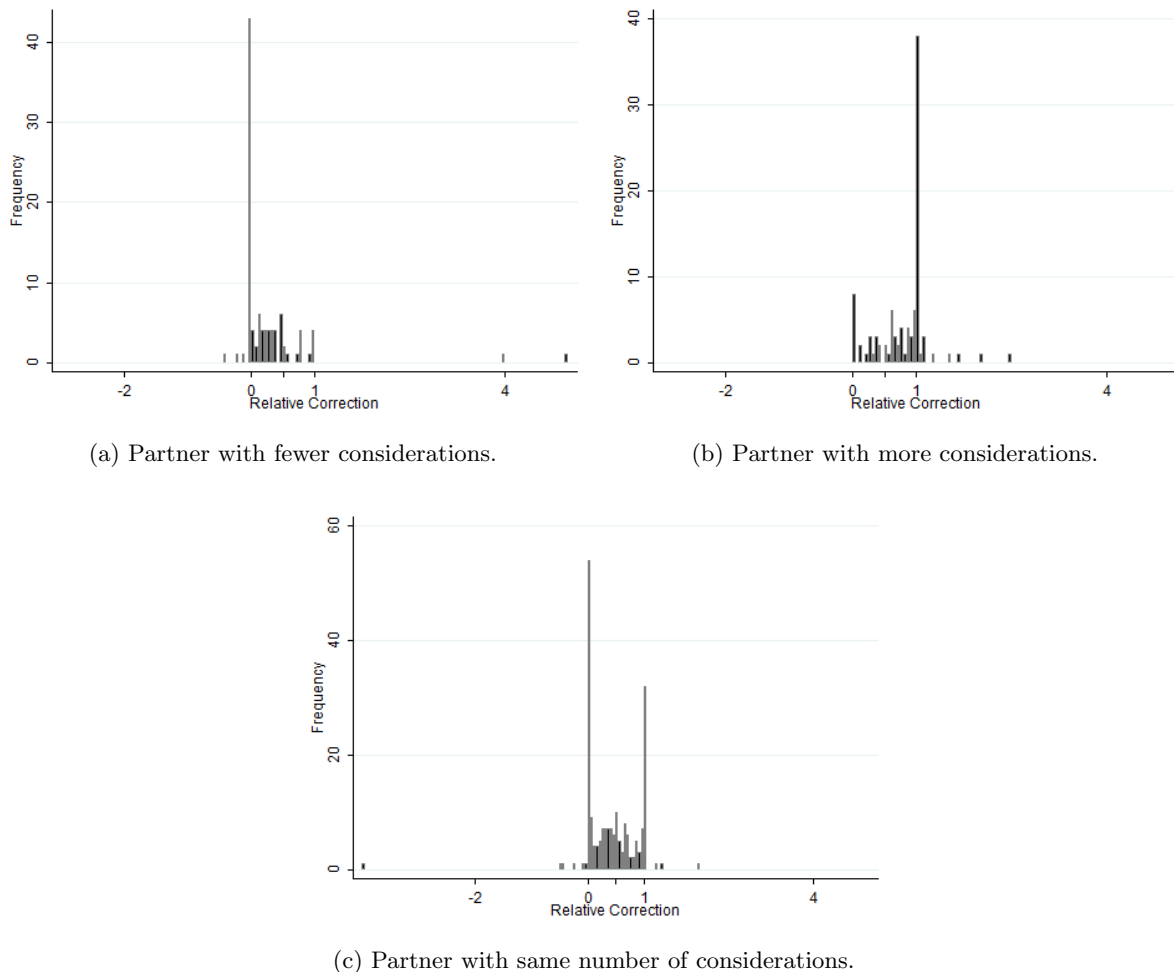


Figure 1: Histograms of the relative correction by the sign of the relative number of informative considerations.

Figure 1 shows that the relative number of informative considerations of the partner is predicting whether the relative correction is predominantly 0 or 1. Having a partner that presents fewer informative considerations (figure 1a) leads to significantly less correction than having a partner that presents more informative considerations (figure 1b, two sample Wilcoxon rank sum test for equality of distributions: $p < 0.0001$). In between lie the relative corrections that follow from a partner with the same number of informative considerations (figure 1c).

To investigate the relevance of the given range of years, the data is divided by the relative precision of the guess. Figure 2 shows the resulting distributions and looks very similar to figure 1. A partner with a less precise guess (figure 2a) induces significantly less correction than a partner with a more precise guess (figure 2b, $p < 0.0001$). In between lie the relative corrections that follow from a partner with a similarly precise guess (figure 2c).¹⁰

¹⁰Similar results are obtained when the difference in precision or the number of considerations is substituted by the difference in number of characters or lines in the message. Those characteristics are strongly correlated

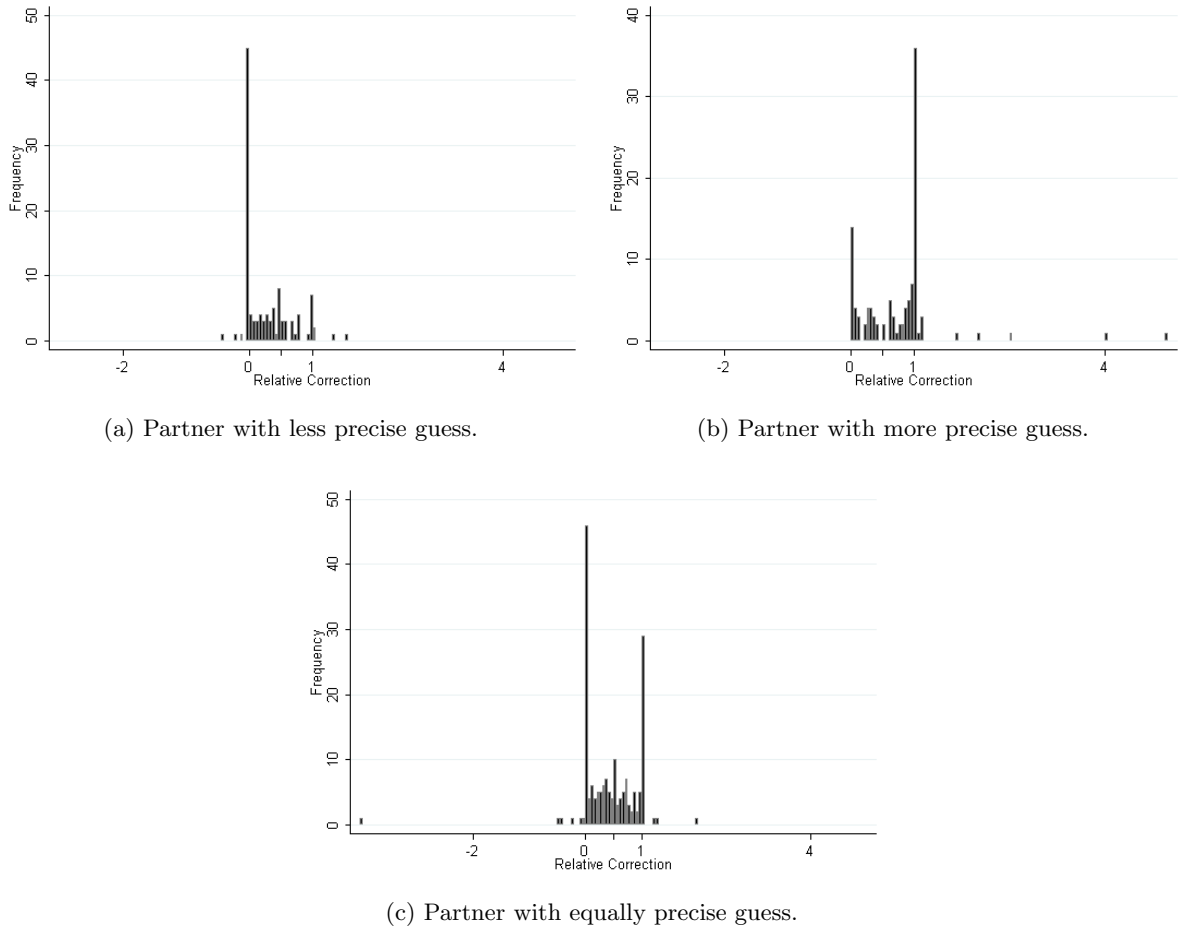


Figure 2: Histograms of the relative correction by the relative precision of the guess.

Showing a regression of differences in all message characteristics on the relative correction, table 5 indicates that the two analyzed characteristics predict the persuasiveness of a message. The number of uninformative considerations or the presentation of credentials have no significant impact on the correction.

More generally, these results suggest that teams aggregate information through individual team members who identify new and more precise considerations and change their decision accordingly. The next section investigates whether such persuasion is observed in teams faced with strategic settings.

7. Results for the beauty contest game

In the beauty contest game, the teams choose a number between 0 and 100 and the one closest to $2/3$ of the average wins a prize. The following example gives an idea of the data obtained in this game.

with the data from the classification.

Table 5: OLS regression of relative correction on message characteristics.

Variables	Rel. correction
Informational consideration difference	0.11** (0.053)
Neutral consideration difference	-0.026 (0.074)
Precision difference	0.00010** (0.000041)
Credential difference	0.11 (0.11)
Constant	0.48*** (0.027)
N	400
R ²	0.142

Notes: Robust standard errors in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% level.

Subject 35 (team 18, suggested decision 77, final decision 27):

i do not have a suggested decision as it is not based on any logic .. i cannot predict what the other teams will play and cannot figure out how to work out what would be the best decision to make. if you have any better ideas, i am willing to take your advice if you dont have any better ideas, then i guess we will just be lucky..or not..

Subject 36 (team 18, suggested decision 23, final decision 25):

the problem is that we have to be BELOW the assumed average - which is tricky cos how should we know? lets think about what's going on in ppl's minds: they're all facing the same question, so i reckon they'll go for rather small numbers, at least below 50 so lets try being below that as well and do something like 25 or maybe even 20 which would be two thirds of their 30-40

The very strategic argumentation of subject 36 is persuasive and leads subject 35 to choose 27. Also, subject 36 increases his decision after receiving a suggested decision of 77.

7.1. The effects of communication

As before, the effectiveness of the communication can be judged by comparing suggested and final decisions. Table 6 shows that the decisions' mean and median are lower after the communication.¹¹ This indicates that the group interaction in form of the message has an effect on the individual decisions. It moves decisions towards the Nash equilibrium, in line with results in the literature. Observing the same effect affirms that group decision making can be investigated when players' interaction is limited to one message exchange.

¹¹The distributions are significantly different, $p = 0.038$ (Wilcoxon signed-rank test).

Table 6: Summary statistics of decisions in the beauty contest game.

Variable	Mean	Std. Dev.	Median	Min	Max	N
Suggested decision	43.93	21.14	40	0	100	84
Final decision	39.73	18.75	35	0	100	84

Table 7 shows how the relative position of the suggestions within a team relate to the directions of correction. Since no team has coinciding suggested decisions from the start, disagreement is either positive or negative. By definition, one team contributes one positive and one negative disagreement each, leading to 42 observations in each category. A majority of 34 players correct their decision downwards, while only 21 correct upwards. 29 players stick to their initial proposal.

Table 7: Directions of disagreement and correction in the beauty contest game.

		<i>Disagreement</i>		
		Neg.	Pos.	Total
<i>Correction</i>	Neg.	32	2	34
	0	10	19	29
	Pos.	0	21	21
Total		42	42	84

The reception of a lower proposal – negative disagreement – is more persuasive than the reception of a higher one. After negative disagreement, nobody corrects upwards, while two players correct downwards after a positive disagreement. Among those who do not correct, the majority has experienced a positive disagreement with their partner.¹² Again, to quantify the effect of the communication on a decision more gradually, the “relative correction” gives the correction divided by the disagreement.

7.2. Levels of reasoning

To investigate the causes for the differential reaction to positive and negative disagreement, I turn to the content of the messages in terms of level of reasoning. Table 8 shows the partner matches by the lower bound. On the diagonal, 34 players are matched with a partner of the same level. 25 players are below and above the diagonal, respectively. Of the latter matches, the majority involves one level-0 partner.

The variable “level difference” summarizes the relative sophistication between partners and is defined as the difference between the partner’s lower bound of level of reasoning and the own

¹²If the “truth” was the Nash equilibrium, the “truth wins” norm (Davis, 1992) would require not to see any upward correction and is therefore not met. If the “truth” was the session’s winning number, 22 (4) players are further away in their final decision than in their suggested decision, while 33 (11) are closer (In brackets the number of players receiving a more sophisticated message). The team performance is better, but does not meet the “truth wins” norm.

Table 8: Level- k lower bounds of subjects and their team partners in the beauty contest game.

Lower bound	Partner's lower bound				
	0	1	2	3	Total
0	20	13	7	1	41
1	13	14	4		31
2	7	4			11
3	1				1
Total	41	31	11	1	84

lower bound. If a player is facing a partner one level below herself, this variable takes the value -1 . Table 9 illustrates disagreement and correction and differentiates by the sign of the level difference.¹³

Table 9: Directions of disagreement and correction by the sign of the level difference.

	<i>Dis.</i>				<i>Dis.</i>			
	Neg.	Pos.	Total		Neg.	Pos.	Total	
<i>Corr.</i>	Neg.	2	1	3	Neg.	19	1	20
	0	1	12	13	0	3	0	3
	Pos.	0	9	9	Pos.	0	2	2
	Total	3	22	25	Total	22	3	25

(a) Negative level difference.

(b) Positive level difference.

	<i>Disagreement</i>			
	Neg.	Pos.	Total	
<i>Correction</i>	Neg.	11	0	11
	0	6	7	13
	Pos.	0	10	10
	Total	17	17	34

(c) Zero level difference.

Table 9a shows that disagreement is mostly positive when the partner is of a lower level. The majority of these proposals are not followed, players predominantly do not correct their decision. Nine subjects do correct upwards, but – as will be seen later – only by a small amount. A very different picture is drawn in table 9b. Players facing higher level partners usually disagree negatively with them. Strikingly, 19 out of 22 players correct in the direction of their partner's proposal, while only three maintain their initial proposal. Table 9c shows the middle ground that players occupy whose team partners exhibit the same level of sophistication. If there is a non-zero correction, it occurs in the same direction as the disagreement.

Equality of the distributions of relative correction for negative versus positive level differences

¹³It is possible that the exchange of the message and further individual deliberation occur simultaneously. Individual deliberation, however, cannot explain the observed strong pattern as it occurs independently of the usefulness of the partner's message. Empirically, "heureka"-learning is rarely observed in the beauty contest game (see Costa-Gomes and Crawford, 2006). Hence, the effect of individual deliberation in the time elapsed between the sending of the own suggestion and the reception of the message should be small.

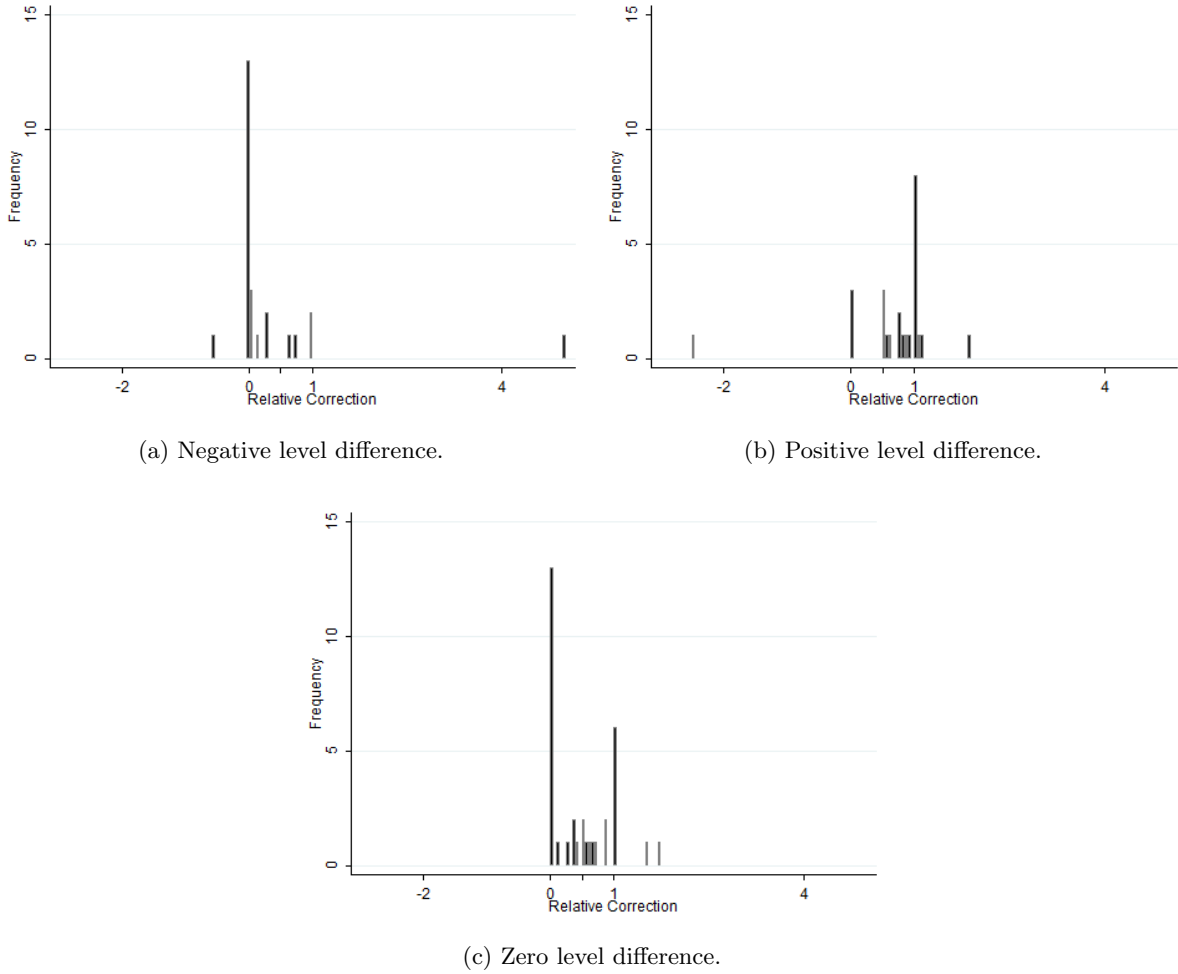


Figure 3: Histograms of the relative correction by the sign of the level difference.

– as depicted in figures 3a and 3b – is rejected by a two sample Wilcoxon rank sum test with a p -value of 0.0005. Comparing the distributions between positive and zero level difference yields a rejection of equality with a p -value of 0.02. The same test with negative and zero level difference rejects equality at the 10% significance level with a p -value of 0.052.

The analysis supports the hypothesis that more sophisticated messages persuade partners more frequently than messages of a lower sophistication. This result illuminates why teams are taking more rational decisions in the sense of actions closer to the Nash equilibrium and why players with a relative correction of 0.5 and higher choose a final decision that is on average 8.5 units closer to the session's winning number than their suggested decision.¹⁴ The team discussion pools arguments and the actions advocated by the more sophisticated team member persuade the less sophisticated player but not vice versa.

¹⁴This result goes in the expected direction but is not statistically significant.

7.3. Level-0 beliefs

To investigate the relation between the persuasiveness of a message and the partners’ differences in the level-0 action or belief, I use the classification data on the level-0 belief which was identified as the non-derived starting point of reasoning in the messages.

Due to the free form of communication, beliefs were not actively elicited. However, in 36 of the 84 messages it was possible to classify the level-0 belief. To complete the data, I make use of the relationship between the suggested decision a , the level of reasoning k , and the level-0 belief b^0 which yields estimates¹⁵ of the level-0 belief:

$$\hat{b}^0 = a \cdot \left(\frac{3}{2}\right)^k .$$

The unbiasedness of this estimator can be checked by comparing its prediction with the data for the 36 classified subjects. Table 10 shows that the estimated data has a similar mean but is as expected more noisy with a higher standard deviation.

Table 10: Level-0 classification and estimation.

Variable	Mean	Std. Dev.	Median	N
Classification	55.26	12.33	50	36
Estimation	53.45	18.53	50	84

I define the “level-0 difference” as the absolute difference between team partners’ level-0 information and split the dataset into observations with a level-0 difference above and below the median of 13. Table 11 shows that a higher difference in the level-0 leads to a higher average relative correction. However, at conventional levels of significance, the equality of the two sample distributions can at that point not be rejected, a higher level-0 difference does not make a message more persuasive per se. Before concluding, a closer look at a possible interaction with the highly persuasive level difference might enable a more differentiated statement.

Table 11: Influence of level-0 difference on relative correction.

Level-0 difference	Mean	Std. Dev.	Median	N	p -value
High	0.560	0.954	0.550	42	0.211
Low	0.420	0.521	0.151	42	

Notes: Statistical comparison of the relative correction in subsamples of players with high vs. low absolute level-0 difference; Wilcoxon rank sum test for equality of distributions.

Table 12 shows the results of the level-0 analysis of table 11 split up into negative, zero, and positive level difference, respectively. It shows that upon reception of a less sophisticated argu-

¹⁵This estimation assumes a degenerate population belief on level $k - 1$, a reasonable assumption given that the majority of players exhibit such a population belief (see Burchardi and Penczynski, 2014).

ment (negative level difference), a high level-0 difference influences the magnitude of correction significantly. High level-0 differences lead to significantly larger relative corrections towards the partner’s proposal. However, a high level-0 difference does not influence the relative correction when a more or equally sophisticated partner is faced. Therefore, the level-0 difference has only an impact when the level of reasoning does not make the argument persuasive. Overall, the analysis identifies circumstances under which level-0 differences can induce a change of decision and therefore be persuasive.

Table 12: Influence of level-0 difference and level difference on relative correction.

Level diff.	Level-0 diff.	Mean	Std. Dev.	Median	N	<i>p</i> -value
Negative	High	0.652	1.306	0.199	14	0.002
	Low	-0.049	0.180	0	11	
Zero	High	0.497	0.412	0.532	14	0.481
	Low	0.455	0.541	0.304	20	
Positive	High	0.530	0.991	0.697	14	0.373
	Low	0.824	0.317	1	11	

Notes: Statistical comparison of the relative correction in subsamples of players with high vs. low absolute level-0 difference and negative, zero and positive level difference; Wilcoxon rank sum test for equality of distributions.

By the Bonferroni correction for multiple hypotheses the critical *p*-value is $0.05/3 = 0.0167$.

Irrespective of the level-0 difference, the mean and median relative corrections for negative level differences are much smaller than those for other level differences. This provides evidence for the larger impact of the level of reasoning than of the level-0 on the persuasiveness of an argument. For a high level-0 difference, these corrections are probably caused by belief updates of higher level players in response to unexpected level-0 information.

This result suggests a second way in which team discussions are fruitful in symmetric strategic settings. The reasoning of team members might inform the belief about opponents’ actions and thus influence the best response. In this way, less sophisticated subjects provide useful inputs in the team discussion. The initial example illustrates this effect.

7.4. OLS regression

In order to recapitulate the analysis of the beauty contest game, I report the OLS regression of level differences and level-0 differences on the relative correction. Regression 1 in table 13 shows the positive and significant impact of the level difference. On average, a difference of one additional level of reasoning results in a relative correction that is 0.19 higher. The impact of the level-0 difference is positive but significant only at the 10%-level. The interaction of level and level-0 difference has – as expected – a negative coefficient since a large level-0 difference is influential for negative level differences only. Regression 2 adds the disagreement variable

and an interaction term with the level difference in order to check whether the mere difference in suggested decisions can explain part of the persuasion. Since negative disagreement leads more frequently to correction, the coefficient is negative, but not significantly different from 0. In regression 3, the variables that result from the messages are dropped and disagreement becomes significant, but it does not explain the relative correction better than the level and level-0 differences. These results suggest that the underlying reasoning of a decision contributes more to persuasion than the mere difference in suggested decisions so that the effect of persuasion indeed results from the communicated arguments rather than the difference in suggestions. This is in line with the fact that very short messages which cannot contain substantial reasoning are usually not followed by a change in the decision. For example, 10 out of 14 players do not correct their decision upon reception of a message of less than 40 characters. Consequently, removing the possibility of a written message would probably lead to much less persuasion.

The impact of the message could also result from its length, reflecting the effort a team partner exerted, independent of its content. Regression 4, therefore, includes the difference in the number of message characters between partner and the player. A longer message of the team partner leads to higher relative correction, but the influence is only significant when the message content variables are excluded (regression 5).

8. Results for the hide and seek game

One might argue that the persuasiveness of higher level arguments is specific to the beauty contest game. A brief and useful robustness check is therefore to see whether similar effects occur in a very different game with a unique mixed strategy Nash equilibrium, no dominant strategies and multiple candidate level-0 beliefs.

In the hide and seek game, hiders hide a treasure and seekers seek it at one of four locations. Whoever holds the treasure at the end, wins a prize. The locations are labelled “ABAA.” For unambiguous reference in this analysis, I index the locations by their positions “ $A_1B_2A_3A_4$.” Using the same communication protocol as before, this experiment features teams of two players that are assigned one of the two roles and matched with a team of the opposite role.

The classification of the written messages provides information about individuals’ level of reasoning analogous to the beauty contest game. The following communication of a team of seekers gives an example of how persuasion might work in this game.

Subject 3 (team 2, suggested decision A_3 , final decision A_4):

for now it's an absolute guess, I would choose the third box

Subject 4 (team 2, suggested decision A_4 , final decision A_1):

the hider may think most people may look for the item in the middle two, thus they

Table 13: OLS regression results, regressing relative correction on various independent variables.

Variables	(1) Rel. corr.	(2) Rel. corr.	(3) Rel. corr.	(4) Rel. corr.	(5) Rel. corr.
Level diff.	0.19*** (0.061)	0.17** (0.077)		0.16** (0.064)	
Level-0 diff.	0.0041* (0.0021)	0.0046** (0.0023)		0.0032 (0.0023)	
Level diff. \times Level-0 diff.	-0.0033* (0.0019)	-0.0058* (0.0034)		-0.0030 (0.0021)	
Disagreement		-0.0038 (0.0035)	-0.0039** (0.0017)		
Level diff. \times Disagreement		0.0012 (0.0015)	0.00067 (0.0016)		
Character diff.				0.00027 (0.00032)	0.00059* (0.00032)
Level diff. \times Character diff.				-0.00042 (0.00028)	-0.00050* (0.00028)
Constant	0.41*** (0.097)	0.42*** (0.11)	0.50*** (0.11)	0.47*** (0.11)	0.55*** (0.090)
Observations	84	84	84	84	84
R-squared	0.057	0.073	0.029	0.072	0.043
Adjusted R^2	0.022	0.014	0.005	0.013	0.019

Notes: Robust standard errors in brackets.

***, **, and * denote significance at the 1%, 5%, and 10% level.

will hide the treasure in the first or last, so i choose the last one

In this brief exchange, subject 4 presents a scenario of hidiers' play and describes a best response. Subject 3 reports to simply "guess" a location and ultimately follows his partner's suggestion.

The exposition of the analysis is influenced by the fact that the action space is a set of only four elements and the unique Nash equilibrium is in mixed strategies. Consequently, observed changes in decisions will be less nuanced and cannot be interpreted as moves towards or away from the Nash equilibrium. Comparing decisions can detect disagreement and correction, but the direction and magnitude of changes in decisions are not meaningful. In other words, it only matters whether subjects propose the same action (disagreement = 0) or not (1) and whether they adopt exactly the decision of their partners (adoption = 1) or not (0).

Table 14 shows disagreement and adoption for 114 observations. The 16 subjects that agreed from the start maintained their position, thereby "adopting" the partner's position. Of 98 subjects that disagreed in their suggested decision, 25 are persuaded by their team partner to adopt exactly her suggested decision.

Table 15 summarizes the partner matches by the level- k lower bound. The diagonal shows that roughly one third of the players are matched with a player of the same level of reasoning,

Table 14: Directions of disagreement and adoption of the partner’s suggested decision.

		<i>Disagreement</i>		
		0	1	Total
<i>Adoption</i>	0	0	73	73
	1	16	25	41
Total		16	98	114

roughly the same as in the beauty contest game. Matches of different level players predominantly involve one player that is of level-0.

Table 15: Level- k lower bounds of subjects and their team partners in the hide and seek game.

Lower bound	Partner’s lower bound					Total
	0	1	2	3	4	
0	24	19	11			54
1	19	12	6	1		38
2	11	6	2		1	20
3		1				1
4			1			1
Total	54	38	20	1	1	114

Table 16 differentiates the data on disagreement and adoption by the sign of the level difference between the team partners. A significant difference in adoption after disagreement can be observed between players of negative and positive level difference. 8% adopt the suggested decision of a lower level partner and 38% adopt the one of a higher level partner (Fisher exact test, $p = 0.002$). These results strengthen the conclusion that the level of reasoning is an important predictor of persuasion.

Because of the nature of the action space, two players’ level-0 beliefs can only be the same or different, a more nuanced comparison is not possible. This game is therefore not well-suited to investigate the level-0 influence on persuasion. The classification data indicates that only in one team both players hold the same level-0 belief from the start. The level-0 belief or action – each player’s reaction to the framing – appears to be more controversial than in the beauty contest game and possibly interferes with the persuasiveness of the level of reasoning. By virtue of this, the hide and seek game provides a strong test whose result consequently emphasizes the relevance of the level of sophistication for persuasion.

9. Discussion

The messages in this study developed naturally in the course of the experiment and were not controlled or manipulated in any way. The possible distinction between the effects of proposal and message is also left for future work. These decisions follow from the objective to study

Table 16: Directions of disagreement and adoption by the sign of the level difference.

		<i>Dis.</i>		
		0	1	Total
<i>Ad.</i>	0	0	32	32
	1	3	3	6
Total		3	35	38

(a) Negative level difference.

		<i>Dis.</i>		
		0	1	Total
<i>Ad.</i>	0	0	20	20
	1	3	15	18
Total		3	35	38

(b) Positive level difference.

		<i>Disagreement</i>		
		0	1	Total
<i>Adoption</i>	0	0	21	21
	1	10	7	17
Total		10	28	38

(c) Zero level difference.

aspects of the *typical* team interaction as it has been implemented in many other experimental studies. It turned out that there were fewer subjects of a level of 2 or beyond than required for a study of the persuasiveness of more sophisticated messages. The scope of the study is thus limited to a standard subject pool, which importantly does not invalidate the relevance of the observed mechanisms for typical team interactions in such settings. An investigation of more sophisticated interaction is a straightforward extension of the proposed methodology which, however, would render results less comparable with other team experiments.

With the data at hand, one cannot distinguish whether the persuasion works via a true understanding of the received argument or due to the authority that a complex and elaborate argument entails, an important question for jury decisions (see Cooper, Bennett and Sukel, 1996). Again, this point can be addressed in a similar setting by varying the content of messages in a controlled fashion.

It has been argued that part of the persuasion might be due to the fact that team members feel obliged to conform with their team partner. While this might lead to more frequent corrections than without the desire of conformity, it does not account for the asymmetry in correction between more and less sophisticated team members. Plain conformity cannot achieve this important coordination in the direction of better founded decisions.

In terms of methodology, the results add another argument for the reliability and replicability of the classification of messages that are obtained in this kind of intra-team communication (Burchardi and Penczynski, 2014). A first argument is that the procedure with separate classifications by two RAs is in accordance with standard procedures of protocol analysis (Ericsson and Simon, 1984). Second, the reliability can be demonstrated by similar classifications of multiple RAs, as shown in table 2. Third now, the classified levels of reasoning and level-0 information are predictive of changes in behavior of the subjects in the experiments. Consequently, there is

something about the messages which causes subjects to change their behavior and at the same time causes classifiers to identify a certain, higher level of reasoning. In this sense, the subjects in the study themselves provide a robustness check of the classification.

Finally, I want to note that the communication protocol in the experiment is a special case of the Delphi method, a group discussion protocol introduced by Dalkey and Helmer (1963). In multiple rounds, experts give individual opinions of, say, the growth of the economy next year. After each round a neutral facilitator distributes anonymous feedback on all others' estimates and justifications. This method has so far been exclusively used as a tool to aggregate expert opinions in an unbiased way, but this study shows its potential to investigate individual and team reasoning.

10. Conclusion

This paper investigates persuasion in the context of team decision making and studies the influence of verbal arguments on individual's decisions in knowledge-related and strategic settings. The study gives concrete insights in the mechanism behind the widely observed team superiority in decision making. An important key to these insights is the experimental design that enables the tracking of individual arguments and decision changes.

In knowledge questions which ask for the year of a historic event, informative considerations and the precision of the guess make a message persuasive in that receivers change their decision more frequently.

Two effects of persuasion can be identified in strategic thinking. First, arguments that exhibit a relatively higher level of reasoning are more persuasive than lower level arguments. It is shown that decisions are more frequently revised in the direction of the partner's proposal when more sophisticated arguments are received. For teams, this implies that the more sophisticated members have a larger influence on the team's decision, rendering the team more sophisticated.

Second, being confronted with unexpected decisions of less sophisticated team members induces higher level players to revise their decision in the direction of their partners' suggestion. This effect can be interpreted as an update of players' beliefs about the behavior of their opponents.

The results show how individuals give up their initial decision when confronted with more sophisticated arguments and maintain their initial decision otherwise. Such a selection process is an important mechanism to improve decisions when arguments are pooled and team members are heterogeneous with respect to individual knowledge or strategic sophistication. While team decision making processes are certainly influenced positively and negatively by many factors, my experimental study illuminates how teams aggregate information and pool strategic sophistication to make better decisions.

References

- Bacharach, Michael and Dale O. Stahl**, “Variable-Frame Level-n Theory,” *Games and Economic Behavior*, August 2000, *32* (2), 220–246.
- Blinder, Alan S. and John Morgan**, “Are Two Heads Better than One? Monetary Policy by Committee,” *Journal of Money, Credit and Banking*, 2005, *37* (5), pp. 789–811.
- Bornstein, Gary and Ilan Yaniv**, “Individual and Group Behavior in the Ultimatum Game: Are Groups More ‘Rational’ Players?,” *Experimental Economics*, 1998, *1* (1), 101–108.
- Bosch-Domènech, Antoni, José G. Montalvo, Rosemarie Nagel, and Albert Satorra**, “One, Two, (Three), Infinity, . . . : Newspaper and Lab Beauty-Contest Experiments,” *American Economic Review*, December 2002, *92* (5), 1687–1701.
- Bosman, Ronald, Heike Hennig-Schmidt, and Frans Van Winden**, “Exploring group decision making in a power-to-take experiment,” *Experimental Economics*, 2006, *9* (1), 35–51.
- Burchardi, Konrad B. and Stefan P. Penczynski**, “Out of your mind: Eliciting individual reasoning in one shot games,” *Games and Economic Behavior*, 2014, *84* (0), 39 – 57.
- Burnstein, Eugene, Amiram Vinokur, and Yaacov Trope**, “Interpersonal comparison versus persuasive argumentation: A more direct test of alternative explanations for group-induced shifts in individual choice,” *Journal of Experimental Social Psychology*, 1973, *9* (3), 236–245.
- Camerer, Colin F., Teck-Hua Ho, and Juin-Kuan Chong**, “A Cognitive Hierarchy Model of Games,” *The Quarterly Journal of Economics*, August 2004, *119* (3), 861–898.
- Casari, Marco, Jingjing Zhang, and Christine Jackson**, “When Do Groups Perform Better than Individuals? An Acquiring a Company Experiment,” Technical Report, University of Bologna 2012.
- Çelen, Bogaçhan, Shachar Kariv, and Andrew Schotter**, “An experimental test of advice and social learning,” *Management Science*, 2010, *56* (10), 1687–1701.
- Charness, Gary and Matthew O. Jackson**, “Group play in games and the role of consent in network formation,” *Journal of Economic Theory*, 2007, *136* (1), 417–445.
- **and Matthias Sutter**, “Groups make better self-interested decisions,” *The Journal of Economic Perspectives*, 2012, pp. 157–176.
- Cooper, David J. and John H. Kagel**, “Are Two Heads Better than One? Team versus Individual Play in Signaling Games,” *American Economic Review*, June 2005, *95* (3), 477–509.

- Cooper, Joel, Elizabeth A. Bennett, and Holly L. Sukel**, “Complex scientific testimony: How do jurors make decisions?,” *Law and Human Behavior*, 1996, 20, 379–394.
- Costa-Gomes, Miguel A. and Vincent P. Crawford**, “Cognition and Behavior in Two-Person Guessing Games: An Experimental Study,” *American Economic Review*, December 2006, 96 (5), 1737–1768.
- Crawford, Vincent P. and Joel Sobel**, “Strategic Information Transmission,” *Econometrica*, November 1982, 50 (6), 1431–51.
- Dalkey, Norman and Olaf Helmer**, “An Experimental Application of the Delphi Method to the Use of Experts,” *Management Science*, 1963, 9 (3), pp. 458–467.
- Davis, James H.**, “Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples 1950–1990,” *Organizational Behavior and Human Decision Processes*, June 1992, 52 (1), 3–38.
- DeMarzo, Peter M., Dimitri Vayanos, and Jeffrey Zwiebel**, “Persuasion Bias, Social Influence, And Unidimensional Opinions,” *The Quarterly Journal of Economics*, August 2003, 118 (3), 909–968.
- Ericsson, K.A. and H.A. Simon**, *Protocol Analysis: Verbal Reports as Data*, Cambridge, Mass.: MIT Press, 1984.
- Fischbacher, Urs**, “z-Tree: Zurich toolbox for ready-made economic experiments,” *Experimental Economics*, June 2007, 10 (2), 171–178.
- Glazer, Jacob and Ariel Rubinstein**, “On Optimal Rules of Persuasion,” *Econometrica*, 2004, 72 (6), 1715–1736.
- and —, “A Model of Persuasion with Boundedly Rational Agents,” *Journal of Political Economy*, 2012, 120 (6), pp. 1057–1082.
- Hastie, Reid**, “Experimental Evidence on Group Accuracy,” in Bernard Grofman and Guillermo Owen, eds., *Information Pooling and Group Decision Making: Proceedings of the Second University of California, Irvine, Conference on Political Economy*, Vol. 2, JAI Press Inc., 1983, pp. 129–157.
- Hill, Gayle W.**, “Group versus individual performance: Are N+1 heads better than one?,” *Psychological Bulletin*, May 1982, 91 (3), 517–539.
- Kamenica, Emir and Matthew Gentzkow**, “Bayesian Persuasion,” *American Economic Review*, 2011, 101 (6), 2590–2615.

- Kocher, Martin G. and Matthias Sutter**, “The Decision Maker Matters: Individual versus Group Behaviour in Experimental ‘Beauty-Contest’ Games,” *The Economic Journal*, January 2005, *115* (500), 200–223.
- Kugler, Tamar, Edgar E Kausel, and Martin G Kocher**, “Are groups more rational than individuals? A review of interactive decision making in groups,” *Wiley Interdisciplinary Reviews: Cognitive Science*, 2012, *3* (4), 471–482.
- Laughlin, Patrick R. and Alan L. Ellis**, “Demonstrability and social combination processes on mathematical intellectual tasks,” *Journal of Experimental Social Psychology*, 1986, *22* (3), 177 – 189.
- Levine, John M. and Richard L. Moreland**, *Small groups: key readings* Key readings in social psychology, New York, NY: Psychology Press, 2006.
- Lipman, Barton L.**, “Decision Theory without Logical Omniscience: Toward an Axiomatic Framework for Bounded Rationality,” *Review of Economic Studies*, April 1999, *66* (2), 339–61.
- Luhan, Wolfgang, Martin Kocher, and Matthias Sutter**, “Group polarization in the team dictator game reconsidered,” *Experimental Economics*, 2009, *12*, 26–41. 10.1007/s10683-007-9188-7.
- Milgrom, Paul and John Roberts**, “Relying on the Information of Interested Parties,” *The RAND Journal of Economics*, 1986, *17* (1), pp. 18–32.
- Modica, Salvatore and Aldo Rustichini**, “Awareness and partitional information structures,” *Theory and Decision*, 1994, *37*, 107–124.
- Mullainathan, Sendhil, Joshua Schwartzstein, and Andrei Shleifer**, “Coarse Thinking and Persuasion,” *The Quarterly Journal of Economics*, May 2008, *123* (2), 577–619.
- Nagel, Rosemarie**, “Unraveling in Guessing Games: An Experimental Study,” *American Economic Review*, December 1995, *85* (5), 1313–1326.
- Penczynski, Stefan P.**, “Strategic Thinking: The Influence of the Game,” Working Paper, University of Mannheim January 2014.
- Rubinstein, Ariel**, *Modeling Bounded Rationality*, Sage Publications, 1998.
- **and Amos Tversky**, “Naive Strategies in Zero-Sum Games,” Working Paper 17-93, The Sackler Institute of Economic Studies 1993.
- Schotter, Andrew**, “Decision Making with Naive Advice,” *American Economic Review*, May 2003, *93* (2), 196–201.

Sheremeta, Roman M and Jingjing Zhang, “Can groups solve the problem of over-bidding in contests?,” *Social Choice and Welfare*, 2010, *35* (2), 175–197.

Stahl, Dale O. and Paul W. Wilson, “On Players’ Models of Other Players: Theory and Experimental Evidence,” *Games and Economic Behavior*, July 1995, *10* (1), 218–254.

Surowiecki, James, *The Wisdom of Crowds*, Anchor, 2005.

Thorndike, Robert L., “The Effect of Discussion upon the Correctness of Group Decisions, when the Factor of Majority Influence is Allowed For,” *The Journal of Social Psychology*, 1938, *9* (3), 343–362.

A. Experiment and Classification Instructions

NOT FOR PUBLICATION

Introduction

This appendix reproduces all instructions shown to the participants in the paper's experiments. The beauty contest and hide and seek games were conducted with Konrad B. Burchardi in the experimental laboratory of Royal Holloway on the 29th of May 2008 and the 28th of November 2008. The trivia sessions were conducted in the experimental laboratory of Cornell University in March and May 2011.

Furthermore, the appendix reproduces the instructions which were provided to the research assistants in order to explain how to classify the written accounts of reasoning and the messages in the trivia questions.

Experiment Strategic Games

The experiment instructions were distributed sequentially before the Test Period, Part I and Part II and read aloud by the experimenter.

Welcome to the experiment!

Introduction

You are about to participate in an experiment in team decision making. The experiment is funded by the Michio Morishima fund, the London School of Economics and the German Society of Experimental Economic Research. Please follow the instructions carefully.

In addition to the participation fee of £5, you may earn a considerable additional amount of money. Your decisions and the decisions of the other participants determine the additional amount. You will be instructed in detail how your earnings depend on your and the others' decisions. All that you earn is yours to keep, and will be paid to you in private, in cash, after today's session.

It is important to us that you remain silent and do not look at other people's screens. If you have any questions or need assistance of any kind, please raise your hand, and an experimenter will come to you. If you talk, exclaim out loud, etc., you will be asked to leave. Thank you.

Since this is a team experiment, you will at various times be matched randomly with another participant in this room, to form a team that plays as one entity. Your team's earnings will be shared equally between you and your team partner.

The experiment consists of two parts (**Part I** and **Part II**) which are independent of each other and feature different tasks. Part I consists of three rounds and Part II consists of four rounds. However, the way you interact as a team to take decisions will be the same throughout the two parts.

Now, let us explain how your *Team's Action* is determined. In fact, both your team partner and you will enter a *Final Decision* individually and the computer will choose randomly which one of your two final decisions counts as your team's action. The probability that your team partner's final decision is chosen is equal to the probability that your final decision will be chosen (i.e. your chances are 50:50). However, you have the possibility to influence your partner's final decision in the following way: Before you enter your final decision, you can propose to your partner a *Suggested Decision* and send him one and only one text *Message*. *Note that this message is your only chance to convince your partner of the reasoning behind your suggested decision. Therefore, use the message to explain your suggested decision to your team partner.* After you finish entering your suggested decision and your message, these will be shown to your team partner. She/he will then make her/his final decision. Similarly, you will receive your partner's suggested decision and message. You will then make your final decision. As outlined above, once you both enter your final decision, the computer chooses randomly one of your final decisions as your team's action.

If you have any questions at this point, please raise your hand. In order for you to get familiar with the messaging system, you will now try it out in a **Test Period**. Please turn the page for further instructions.

Test period

A participant in this room is now randomly chosen to be your team partner. The **Test Period** has two rounds, with one question to answer in each round. Since this is only a test, your earnings will not depend on any decision taken now. In both test rounds you will need to answer a question about the year of an historic event. The team that is closest to the correct year wins.

As described, you will be able to send one *Suggested Decision* with your proposed year and an explaining *Message*. After having read your partner's suggested decision and message, you will enter your *Final Decision*. As described earlier, either your or your partner's final decision will be chosen randomly to be your *Team's Action*.

The messenger allows *Messages* of any size. However, you have to enter the message line by line since the input space is only one line. Within this line you can delete by using the usual "Backspace" button of your keyboard. By pressing "Enter" on the keyboard, you add the written sentence to the message. Please note that only added sentences will be sent and seen by your partner. *The words in the blue input line will not be sent.* You can always delete previously added sentences by clicking the "Clear Input" button. The number of lines you send is not limited. You can therefore send messages of any length. You finally send the message to your partner by clicking the "Send Message" button.

When you are ready, please click the "Ready" button to start the **Test Period**.

Start Part I

You are about to start Part I of the experiment. You are now randomly matched with a new team partner. For each of the next three rounds you will be matched with a new team partner, i.e. in each of the following rounds you will play with a different person.

In each round, once all teams' actions have been taken, the computer will let you know which of your final decisions has been chosen randomly as your team's action. It will also let you know all other teams' decisions, whether your team won the round and your personal earnings. In each round, the winning team earns £20 (£10 per team player).

Then the next round of the game follows. It will feature an identical task, but you will be matched with a new team partner.

Your task is the following:

Your team and all other teams will take their *Team's Action* by choosing a number between 0 and 100. 0 and 100 are also possible. Only whole numbers will be accepted. From all teams' actions, the computer will calculate the *Average*. Two thirds ($\frac{2}{3}$) of this average will be your *target* number. The winning team will be the one that is *closest to two thirds of the average*. If two or more teams are equally close, the prize will be randomly given to one of these teams.

As described earlier, you will send your team partner a *Suggested Decision* and a *Message*. Remember to explain in the message your reasoning behind your suggested decision. (*And note again that the words in the blue input line will not be sent. Press "Enter" to add them to the message.*) After this information is exchanged, both of you enter your *Final Decision*, from which the computer randomly chooses the *Team's Action*.

When you click the "Ready" button, you will start the first round of **Part I** of the experiment.

Start Part II

You are about to start Part II of the experiment. You are now randomly matched with a new partner. For each of the next four rounds you will be matched with a new team partner, i.e. in each of the following rounds you will play with a different person.

In this part of the experiment your team will play against only one other team. In each of the four rounds you play against a different team. From the four rounds, one round is chosen randomly and will be considered for determining the payoff. If your team wins this selected round, your team will earn £10 (£5 per team player). Please note that you will be informed of your opponent's team action of the chosen round at the end of Part II. There will be *no feedback* after the individual rounds.

Your task is the following:

In the beginning, the computer will tell you whether your role throughout Part II is "*Hider*" or "*Seeker*".

If you are *Hider*, your task is to hide an object behind one of four items. In rounds 1 and 2, the object is a *Treasure*. In rounds 3 and 4, the object is a *Mine*. The hider team wins the round if the treasure was not found by the seeker or if the mine was found by the seeker. The seeker does not observe where you hide the object. The seeker will look behind one item in each round, not more and not less.

If you are *Seeker*, your task is to find the treasure in rounds 1 and 2 and to avoid the mine in rounds 3 and 4. The seeker team wins the round if it chooses the particular item behind which the treasure *was* hidden or if it chooses an item behind which the mine *was not* hidden.

Just like in Part I, you can send a *Suggested Decision* and an explaining *Message* to your team partner. (And note again that the words in the blue input line will **not** be sent. Press “Enter” to add them to the message.) From your two *Final Decisions* the computer again chooses the *Team’s Action*.

When you click the “Ready” button, you will start the first round of **Part II** of the experiment.

Experiment Trivia Questions

Welcome to the experiment!

Introduction

You are about to participate in an experiment in team decision making. The experiment is funded by Cornell University. Please follow the instructions carefully.

In addition to the participation fee of \$5, you may earn a considerable additional amount of money. Your decisions determine the additional amount. You will be instructed in detail how your earnings depend on your decisions. All that you earn is yours to keep, and will be paid to you in private, in cash, after today’s session.

It is important to us that you remain silent and do not look at other people’s screens. If you have any questions or need assistance of any kind, please raise your hand, and an experimenter will come to you. If you talk, shout out loud, etc., you will be asked to leave. Thank you.

The experiment consists of a test round and 2 parts. Part I is designed as a warm-up with 3 trivia questions. Part II consists of 2 rounds of 3 decisions each. In all 6 decisions your task will be identical and success is identically rewarded.

Since this is a team experiment, you will be randomly matched with another participant in this room, to form a team that plays as one entity. Your teammate will change every round, so please do not assume content of previous communication to be known by your new partner. The way you interact as a team to take decisions will be the same throughout all rounds.

Now, let me explain how your *Team’s Action* is determined. In fact, both your teammate and you will enter a *Final Decision* individually and the computer will choose randomly which one of your two final decisions counts as your team’s action. The probability that your teammate’s final decision is chosen is equal to the probability that your final decision will be chosen (i. e. your chances are 50:50). However, you have the possibility to influence your partner’s final decision in the following way: Before you enter your final decision, you can propose to your partner a *Suggested Decision* and send him one and only one text *Message*. Note that this message is your only chance to convince your partner of the reasoning behind your suggested decision. Therefore, use the message to explain your suggested decision to your teammate. After you finish entering your suggested decision and your message, these will be shown to your teammate. Simultaneously, you will receive your partner’s suggested decision and message. Both of you will then make your final decision. As outlined above, once you both enter your final decision, the computer chooses randomly one of your final decisions as your team’s action.

If you have any questions at this point, please raise your hand. In order for you to get familiar with the messaging system, you will now try it out in a **Test Period**. Please turn the page for further instructions.

Test period

A participant in this room is now randomly chosen to be your teammate. The **Test Period** has two rounds, with communication in each round. Since this is only a test, your earnings will not depend on anything that happens now. In both test rounds you will need to send and receive pieces of information. The information consists of the answer to a question and one given phrase. After the successful exchange, you will enter the number again. This way, the communication structure is identical to the one in the experiment rounds.

The messenger allows *Messages* of any size. However, you have to enter the message line by line since the input space is only one line. Within this line you can delete by using the usual “Backspace” button of your keyboard. By pressing “Enter” on the keyboard, you add the written sentence to the message. Please note that *only* added sentences will be sent and seen by your partner. *The words in the blue input line will not be sent.* You can always delete previously added sentences by clicking the “Clear Input” button. The number of lines you send is not limited. You can therefore send messages of any length. You *finally* send the message to your partner by clicking the “Send Message” button.

If you have any questions at this point, please raise your hand. When you are ready, please click the “Ready” button to start the **Test Period**.

Experiment - Part I

In **Part I** you are asked to answer three trivia questions. All questions are about the year of an historic event. For each question, the team that is closest to the correct year wins a prize of \$10, \$5 per team member. If more than 1 team are equally close to the correct year, the winner will be randomly determined among them. You will be matched with a different teammate for each question.

As described, you will be able to send one *Suggested Decision* with your proposed year and an explaining *Message*. After having read your partner’s suggested decision and message, you will enter your *Final Decision*. As described earlier, either your or your partner’s final decision will be chosen randomly to be your *Team’s Action*.

The messenger allows *Messages* of any size. However, you have to enter the message line by line since the input space is only one line. Within this line you can delete by using the usual “Backspace” button of your keyboard. By pressing “Enter” on the keyboard, you add the written sentence to the message. Please note that *only* added sentences will be sent and seen by your partner. *The words in the blue input line will not be sent.* You can always delete previously added sentences by clicking the “Clear Input” button. The number of lines you send is not limited. You can therefore send messages of any length. You *finally* send the message to your partner by clicking the “Send Message” button.

If you have any questions at this point, please raise your hand. When you are ready, please click the “Ready” button to start **Part I**.

Experiment - Part II

Part II is not relevant for this study.

Classification Beauty Contest

The self-contained classification instructions were given to the RAs. Part 2 was handed out once the classification of Part 1 was finished.

Part 1

In the following we will describe the classification process for the analysis of our experiment. We, Konrad and Stefan, assume that you are familiar with the level-k model as it has been introduced by Nagel (1995) or represented by Camerer et al. (2004). However, in order to clarify potential questions of terminology, the appendix on page 36 reproduces the main features of the model in the terminology used in this document.

The classification proceeds in two steps, Part 1 and Part 2. You are now provided by us with the transcripts for Part 1. The transcripts differ in the amount of information about the decisions taken. Only in Part 2 will you see the choices of the players that were made.

After your individual classification of each part, you will meet with your co-classifier to reconcile your classification. In this process, try to agree on common classifications if possible and note them in the third sheet. If an agreement is not possible and you keep your initial individual classification, simply note nothing in the third sheet. After you finished this process for Part 1, you will hand in the three sheets and we will provide you with the material for Part 2. If you have questions about the procedure at any point, simply write an email to us and we will clarify any point in a mail to both of you.

For Part 1, follow the instructions of this booklet now. Read them entirely to get an overview and then start the classification. Please read the messages of each player in Period 1 and note for each player the minimum level of reasoning, the level-0 belief mean and the population distribution. Below you find detailed instructions for classifying each player. Please limit yourself to making inferences only from what can clearly be derived from the message stated, i.e. do not try to think about what the player *might have thought*.

IMPORTANT: When you think that the information does not clearly lend itself to any inference, simply do not note any classification. Consequently, do not note anything if no statement has been made! It seems that this time the statements are less clear than in the first round. Please note only those classifications for which you are certain. Make use of the comments space if you are not certain but still want to indicate a feature of the reasoning. Similarly, please comment if the statement exhibits some argument that does not fit the level-k model as we present it here.

Levels: Minimum lower bounds

For the minimum lower bound on the level of reasoning, you should ask yourself: “What is the minimum level of reasoning that this statement clearly exhibits?” Once noted, you should be able to say to yourself: “It seems impossible that the players’ level of reasoning is below this number!”

Here we ask you to be very cautious with the classification, not giving away high levels easily. *Please only write down the highest lower-bound for which you are absolutely certain!* In part 2 you will be asked to classify the maximum lower bounds of the level of reasoning. This will be the time to be generous with the interpretation of the statements.

Level 0 The player does not exhibit any strategic reasoning whatsoever. Different versions of this might be randomly chosen numbers, misunderstanding of the game structure or giving other non-strategic 'reasons' for picking a number, e.g. taste. Important is that no best-responding to the others' play occurs. There could be considerations of what others might play, but without best responding to it. Examples¹⁶: "Let's use 50. This is the average between 0 and 100." "It's random, so let's guess something." "My favourite number is 74."

Level 1 This player best responds to something (calculates 'two thirds'). However, he does not realise that others will be strategic as well. Example: "They will all go for a number of about 50-55. So we should do something like 35."

Level 2 This player not only best responds (calculating 'two thirds'), but also realises that other players best respond as well. At level 2 the question about the extent of strategic reasoning of other players can come up. In the theory, this is reflected as a population belief on levels 0 and 1. Example: "Thinking that others play 60, everybody will play 40. So, we should be more clever and play two thirds of 40." "Some will play just play 90, while others will think and play 60 in response. We should therefore play somewhere between 60 and 40."

Level 3 This player realises that others could be level 2 and reacts by best responding to this as well. Put differently, he realises that others realise that others best respond as well. As for all players above level 1, the extent of strategic reasoning by others is important for level 3 reasoners as well. In addition, they have to ask themselves how the level-2 players think about the distribution of level 1 and level 0 players.

Level 4, 5, ... The process continues to higher levels. More levels of best responses and higher orders of beliefs become relevant.

Level 0 belief mean

If the comment hints toward a value of the mean of the level-0 distribution, then indicate this value as level-0 mean. Remember, the level-0 mean is the starting point of the reasoning. Players of positive level start best responding on this number. Please note a number as level-0 mean only when this number is not logically derived through level reasoning or the like. If an interval is indicated, please note the average of the lower and upper bound. For example, 'I think the others play around 50-60.' can be noted as a mean of 55. If only a qualitative statement is made about the level-0 mean, try to quantify it if possible. Otherwise, please write a short comment that indicates what is written down. Similarly, if a distribution is specified, please comment precisely on the relevant passage.

The literature usually assumes a mean of 50. Be reminded that this is only a common assumption which should not influence your considerations at this point.

¹⁶All examples have been made up for illustrative purposes.

Population belief

The population belief distribution $g_k(h)$ of a player of level k gives the fraction in the population he expects to be of level h . By definition, for level 0 players, the population belief is irrelevant. Level 1 players are defined as believing that all others are level 0. Hence, differences in the population belief distribution can only show up for players who are level 2 or higher. Therefore, we do not expect any statement from you for reasoners below level 2. But even if the level is 2 or higher, be reminded that at points where you think the information does not lend itself to any inference, simply do not note any classification.

We want to distinguish two sorts of population beliefs, distinguished by the degenerateness of the population distribution.

Degenerate Under a degenerate population belief, a player believes that *all* other players reason exactly one level below themselves. A level 5 player believes everybody else to be level 4. Higher order beliefs are also degenerate. So a level 3 player would think that all others (who are believed to be level 2 players) will believe that all others are level 1 and so on. Example: “Thinking that others play 60, everybody will play 40. So, we should be more clever and play two thirds of 40.”¹⁷

Non-degenerate A non-degenerate population belief gives non-zero probability to more than one lower level. In such a case, a level 2 player believes that both level 0 and level 1 subjects are in the game with positive probability. Higher order beliefs could be either degenerate or non-degenerate. An example would be: “Some will play just play 90, while others will think and play 60 in response. We should therefore play somewhere between 60 and 40.” (In the unlikely case, that the population belief is of the ‘degenerate’ type but some higher order beliefs are not, please make a note.)¹⁸

Model and terminology

The level- k model of bounded rationality assumes that players only think through a certain number (k) of best responses. The model has four main ingredients:

Population distribution This distribution on \mathbb{N}_0 reflects the proportion of types with a certain level k .

Level-0 distribution By definition, a level-0 player does not best respond. Hence, his actions are random to the game and distributed over the action space, which in our case is $\mathcal{A} = \{0, 1, 2, \dots, 99, 100\}$.

Level-0 belief In the model, players with $k > 0$ best respond to what they believe the level-0 players play. Their level-0 belief might not be consistent with the level-0 distribution. For best responding, all that matters of the level-0 belief is the mean, which lies in $[0, 100]$. It is frequently assumed that the level-0 distribution and the level-0 belief are consistent, but for the classification this is irrelevant.

Population belief Players do not expect other players to be of the same or a higher level of reasoning. For a level- k player, the population belief is therefore defined on the set of levels strictly below k . It follows that level-0 players have no defined belief, level-1 players have a trivial belief with

¹⁷Nagel (1995) proposed such a population belief.

¹⁸Camerer et al. (2004) proposed a truncated Poisson distribution as population belief distribution.

full probability mass on $\{0\}$, level-2 players have a well defined belief on $\{\{0\}, \{1\}\}$. From level 3 higher order beliefs are relevant as level-3 players have to form a belief about level-2's beliefs.

Part 2

The classification proceeds in two steps, Part 1 and Part 2. You are now provided by us with the transcripts for Part 2. You can now see the choices of the players that were made.

Please consider the information on each player in Period 1 and note for each player the maximum lower bound of level of reasoning and whether the equilibrium has been identified and whether dominance reasoning has been applied in the excel sheet provided. Below you find detailed instructions for classifying each player. Please limit yourself to making inferences only from what can clearly be derived from the message and the action data, i.e. do not try to think about what the player might have thought.

Be reminded that when you think that the information does not clearly lend itself to any inference, simply do not note any classification. Consequently, do not note anything if no statement has been made! It seems that this time the statements are less clear than in the first round. Please note only those classifications for which you are certain. Make use of the comments space if you are not certain but still want to indicate a feature of the reasoning. Similarly, please comment if the statement exhibits some argument that does not fit the level-k model as we present it here.

Levels: Upper bounds

The upper bounds should give the maximum level of reasoning that could be interpreted into the statement. Therefore, you should ask yourself: "What is the highest level of reasoning that can be underlying this statement?" Once noted, you should be able to say: "Although maybe not clearly communicated, this statement could be an expression of this level. If the player reasoned higher than this number, this was not expressed in the statement!"

Please refer to the level characterisations in Part 1 of the instructions.

Type: Equilibrium identification

With this dummy, you indicate whether the player realised that the unique equilibrium is 0. For this he has to *mention* the equilibrium action 0. It is not enough to describe a process of downward convergence. The equilibrium might be mentioned anywhere in the statement, so it is irrelevant whether he stops reasoning when he found the equilibrium strategies or whether finds further arguments not to play 0. People will not necessarily use the word 'equilibrium', but they might describe that 'theoretically everybody should play 0' or that 'the process will be going down to 0'.

Set the dummy to 0 if the equilibrium was not identified and to 1 if it was.

Type: Dominance reasoning

With this dummy, you indicate whether the reasoning applied the concept of dominance for the explanation. This is defined as involving iterative deletion of dominated strategies and randomly playing one of the remaining actions or best responding to a distribution over the partner's remaining action space.

People will not necessarily use the word ‘dominance’, but they might describe that ‘playing above 66 makes no sense’. Note that ‘everybody plays on average 50 so I should not play higher than 34’ is *not* a dominance reasoning, because it rules out strategies based on a distribution on the full action space. Dominance reasoning rules out first and plays a best response then.

Set the dummy to 1 if dominance was used in the argument and to 0 if it was not.

Classification Hide and Seek

Introduction

In the following I will describe the classification process for the analysis of the experiment. It is assumed that you are familiar with the level- k model as it has been introduced by Nagel (1995) or represented by Camerer et al. (2004). The model here is extended to incorporate salience in the level-0 belief according to Bacharach and Stahl (2000). In order to clarify potential questions of terminology and introduce the main features of the model, appendix A reproduces the main features of the model in the terminology used in this document.

After your individual classification, you will meet with your co-classifier to reconcile your classification. In this process, try to agree on common classifications if possible and note them in the third sheet provided. If an agreement is not possible and both of you keep your initial individual classification, simply note nothing in the third sheet. If you have questions about the procedure at any point, please write an email to me and I will clarify any point in an email to both of you.

Please read this document and the instructions for the experiment entirely in order to get an overview and then start the classification based on the player’s sent message and action proposal. Note that the framing of the four possible locations is ‘ABAA’ in periods 1 and 3 and ‘1234’ in periods 2 and 4. A player can be of two types, hider or seeker. It is useful to go through the process first for all hidere and then for all seekers. This way, you keep the perspective of one type of player and do not get confused.

Please read the messages of each player, taking into account his action, and note for each player every possible level of reasoning. These should lie in an interval between the lower and upper bound of the level reasoning, as specified later in detail. Below you find detailed instructions for classifying each player. It is important that you limit yourself to making inferences only from what can clearly be derived from the message stated, i.e. do not try to think about what the player *might* have thought.

IMPORTANT: When you think that the information does not clearly lend itself to any inference, simply do not note any classification. Consequently, do not note anything if no statement has been made! Please note only those classifications for which you are certain. Make use of the comments space if you are not certain but still want to indicate a feature of the reasoning. Similarly, please comment if the statement exhibits some argument that does not fit the level- k model as I present it here.

Levels

For the lower bound on the level of reasoning, you should ask yourself: “What is the minimum level of reasoning that this statement clearly exhibits?” Once noted, you should be able to say to yourself: “It

seems impossible that the players' level of reasoning is below this number!" Here I ask you to be very cautious with the classification, not giving away high levels easily.

The upper bounds should give the maximum level of reasoning that could be interpreted into the statement. Therefore, you should ask yourself: "What is the highest level of reasoning that can be underlying this statement?" Once noted, you should be able to say: "Although maybe not clearly communicated, this statement could be an expression of this level. If the player reasoned higher than this number, this was not expressed in the statement!"

For both lower and upper bound, please refer to the following characterisation of the different levels.

Note that there are two necessary conditions for a player to exhibit a level greater than 0. First, the player has to be responsive to the salience of the games' framing. Secondly, the player has to be strategic in best-responding to his level-0 belief, which is shaped by salience. If he did not react to salience, he would have no reason to chose one over the other object, resulting in random play. Interestingly, this random play is observationally equivalent to equilibrium behaviour. Therefore, the level-0 players can be those that react to salience and do not play strategically, or they can ignore the framing and hence play randomly. As far as possible, we want to distinguish uniform random level-0 play and equilibrium play. However, regarding the level classification, you can classify every random play as level-0 play. The equilibrium play is taken into account by a specific dummy.

Level 0 The player does not exhibit any strategic reasoning whatsoever. Different versions of this might be randomly chosen or purely guessed actions, misunderstanding of the game structure or other non-strategic 'reasons' for picking a location, e.g. by taste or salience. It is important that no best-responding to the other's play occurs. There could be considerations of what others might play, but without best responding to it. Examples¹⁹: "Well, it's a pure guess", "There are no arguments. Simply choose any."

Level 1 This player best responds to some belief (in the treasure game, a hider mismatches the belief of the other's action, a seeker matches his belief of the other's action). However, he does not realise that others will be strategic as well. Example: "They are probably picking B, so we do as well", "Its at the end and people would naturally go for the middle, no?"

Level 2 This player not only shows the basic strategic consideration of playing best response (matching/mismatching), but also realises that other players best respond as well according to the belief they entertain. A level-2 player clearly contemplates how the other player might best respond to his frame. The player plays a best response to this hypothesised consideration. Example: "They may think most will look in the middle two, therefore choosing one in the end. I therefore choose the first one."

Level 3 This player realises that others could be level-2 and reacts by best responding to the associated expected play. Put differently, he realises that others realise that others best respond to their initial belief. Therefore, a level-3 player clearly states that his opponent expects that he (the level-3 player at question) best-responds to a certain belief.

Level 4, 5, ... The process goes on in a similar fashion. The reasoning enters a cycle, in the sense that level-5 will come to the same best response as level-1. Please indicate in the comment section if

¹⁹All examples have been made up for illustrative purposes.

the player reaches this cycling phase and recognises this pattern.

Level-0 belief

If level reasoning is observed in the statement, there has to be a starting point in the argument which states an attraction or aversion to one or more of the locations. This is then not derived by strategic reasons, but is an intuitive reaction to the framing of the locations. Otherwise, level reasoning would not occur.

Please indicate the underlying level-0 belief that is connected with each possible level of reasoning. Note that the level-0 belief of a person reasoning on an odd level, i.e. level 1, 3, 5, etc. is always with respect to how a player of the opposite side intuitively reacts to the framing. The belief of a person reasoning on an even level, i.e. level 2, 4, 6, etc. is always with respect to what the opposite type believes about the own type's intuitive reaction. You will have this in mind and can note it for completeness in the "H/S"-box, but since it follows from your level indication, it is not essential.

Usually, the most information one can get out of the communication is the most and least attractive location respectively. For example, as a seeker, I might communicate that the hider is most attracted to the *B* and then play central *A*. This indicates that the hider was believed to choose central *A* with the (weakly) lowest probability and *B* with highest probability. It follows that it will usually not be possible to rank *all* the locations by their attractiveness.

To reflect the exhibited level-0 beliefs please denote every level-0 belief by ranking the four locations with a 'more attractive than' relation. The locations are coded according to their position as '1', '2', '3', and '4'. In the example of the previous paragraph, the resulting statement is $2 > 14 > 3$, since *B* in the 2nd position is the most attractive location for the hider and central *A*, the '3', the least attractive. Not putting anything between numbers indicates that their level of attractiveness cannot be distinguished, as in this case with the two *A*'s at the beginning (position '1') and at the end (position '4'). Of course, depending on whether you get more or less information out of the communication, your statement can be $14 > 23$ or $2 > 4 > 3 > 1$. This notation should be flexible enough to encode every piece of information on the level-0 belief that is present in the communication.

Imagine a seeker that you classify to be level-1 or level-2. If you determined his level-0 belief to be $2 > 134$ (H) in the case he is level-1, then the level-0 belief should follow to be $134 > 2$ (S) in the case he is level-2.²⁰ For all cases where nothing in the communication speaks against this, feel free only to note the level-0 belief for the first conjectured level. A statement on each possible level's level-0 belief is only necessary if you think it does not follow in this mechanical way.

In order to keep the overview over the best responses that are connected with certain types at certain levels, I will make a small Excel-sheet available, which calculates automatically the best responses as a function of a specified level-0 belief. This will help you to get a feeling for the mapping of communication and action into the parameters.

²⁰This follows because for this last level-0 belief for a seeker, the previous level-0 belief of the hider constitutes a best response. Given the action of the player at hand, this is the way the levels and level-0 beliefs co-move.

Dummy

Equilibrium play In the description of the levels, I said that the level-0 player that does not react to the salience is considered to play randomly due to a lack of arguments for a specific option. Similarly, a player that plays according to the Nash equilibrium will have no argument for or against a specific option, therefore exhibiting the exact same behaviour. In order to distinguish the two where possible, please indicate here whether the player gives convincing arguments for his random play by mentioning that any location is a best response to random play. It is important that the fully random play of the others is considered and used in the best response argument. Put '1' if the player does so. Otherwise, the player will fit the description of level-0 players and the dummy should take the value '0'. Do not tick anything if the player is clearly neither one nor the other type or if the statement does not allow for a classification along these lines.

Concluding remarks

You might have noticed that there will be no classification of the population belief. This is because the opponents as a team are a single entity and a non-degenerate population belief makes only sense in a probabilistic interpretation. Also, the action will always²¹ maximally reflect the mode of the probabilistic population belief, making it observationally equivalent to a degenerate population belief. This is why the population belief has not been discussed in the context of this game in the literature. At the current point of the study I let these details remain in the background. Still, please indicate in a comment if a player exhibits any non-degenerate population belief.

Compared to the Beauty Contest, it might at times be difficult to distinguish what is a level-0 belief and what is derived through level reasoning. Try to stick to what is written down and look for clearly stated arguments of reasoning.

Model and terminology

The level- k model of bounded rationality assumes that players only think through a certain number (k) of best responses.²² The model has four main ingredients:

Population distribution This distribution reflects the proportion of types with a certain level $k \in \mathbb{N}_0 = \{0, 1, 2, 3, 4, 5, \dots\}$.

Level-0 distribution By definition, a level-0 player does not best respond. Hence, his actions are random to the game and distributed randomly over the action space. In our case, the action space is $\mathcal{A} = \{\{1\}, \{2\}, \{3\}, \{4\}\}$ and contains the four possible locations to hide or seek the object. The model incorporates salience by assuming higher probabilities in the level-0 distribution for actions that are salient. In our case, the level-0 distribution would not assign a uniform probability of 0.25 to each possible action, but $p > 0.25$ to the salient one and $q_i < p$ for the remaining actions.

Level-0 belief In the model, the best responses of players with $k > 0$ are anchored in what they believe the level-0 players play. Their level-0 belief might not be consistent with the level-0 distribution. For best responding, all that matters is the expected payoff from choosing a particular location.

²¹I abstract from considerations due to the cycling of behaviour from level 5 onwards.

²²See the paper by Camerer et al. (2004) for a more detailed account of one version of the model.

One would therefore seek (hide) the treasure where the probability is highest (lowest), that the opponent chooses the same location.

Population belief Players do not expect other players to be of the same or a higher level of reasoning. For a level- k player, the population belief is therefore defined on the set of levels strictly below k . It follows that level-0 players have no defined belief, level-1 players have a trivial belief with full probability mass on $\{0\}$, level-2 players have a well defined belief on $\{\{0\}, \{1\}\}$. From level 3 higher order beliefs are relevant as level-3 players have to form a belief about level-2's beliefs.

Classification Trivia Questions

Introduction

In the following I will describe the classification process for the analysis of an experiment. Subjects are deliberating a trivia question that asks them to date a historic event. Their deliberation will be classified along the lines of few criteria introduced below. The aim of the classification is to quantify the informational content of each participant's message.

Follow the instructions of this booklet. Read them entirely to get an overview and then start the classification. The trivia questions are indicated in the booklets that contain the messages. The subjects faced 3 different questions and decided in a team of two players about their action. The winning team was the one closest to the true year and each team member earned \$5.

The questions are all designed to make knowledge of the true year very unlikely, revolving mostly around the year of birth, death or marriage of a very prominent figure of history. Furthermore, they give one piece of information about the person, which is designed to give a first lead for participants to locate the event in time.

Please limit yourself to making inferences only from what can clearly be derived from the message stated.

Pieces of evidence

The first quantifier of the informational content of a message is the number of pieces of evidence. Any fact (apart from that stated in the question) can be such a piece of evidence. For example, the association of the historic figure with renaissance or the dark ages, his involvement in a war, revolution, the period of his prime activity, the number of children, his age at death, etc.

I want to distinguish those pieces of evidence that are used to locate the event in time versus those that do not. For example, indicating that "Gutenberg was active before the Lutheran Reformation in the 1500s" is a **informative** piece of evidence that allows to make the statement that he probably died in the 1400s. However, stating that "Lincoln was lefthanded" would be a **neutral** piece of evidence. I will ask you to count the number of both types of pieces of evidence, respectively.

Precision of guess

The precision of the guess is a number that indicates to how many years player's arguments have limited the possible answers. If the message indicates a range, simply note the magnitude of this range. For

example, ‘I would say that Elvis was born in the 20s or 30s’ would make the range equal to 20.

Some message do not indicate ranges explicitly, but implicitly, talking about beginning of the century, early in the decade etc. In those cases, please use the following ranges:

Range	Phrase about	Example
2000	no indication	<i>Guessing...</i>
100	century	<i>Columbus was in the 1400s.</i>
33	early/beginning/mid/late/end century	<i>I am just guessing early 1700s.</i>
10	decade	<i>Edison was probably born in the 1860s.</i>
3	early/beginning/mid/late/end decade	<i>Edison was probably born in the late 1840s.</i>
1	year	<i>Edison was probably born in 1847.</i>

Combinations of those would be treated analogously. For example, the range for ‘Charlie Chaplin was probably born end of 19th or beginning of 20th century’ would be 66.

Indication of uncertainty

The indication of uncertainty is to capture qualifiers of the player regarding his certainty about his proposal. Two types of uncertainty are distinguished.

Players can state that they are completely clueless, for example, ‘i have no idea.’ Please indicate that with a uncertainty of 2.

Between this cluelessness and no qualifying at all, players might indicate that they are not sure, for example, ‘I’m not sure which year he was born, but I think it was 1460’. Please indicate that with an uncertainty of 1.

If no qualifying comment has been made, please indicate this with 0.

Credential

It is possible that a subject gives a credential that testifies her/his expertise in the area of the question. These could be certain nationalities, majors of study, profession of parents, etc. For example, to say ‘I am from the same country as Copernikus’ alludes to a higher likelihood of knowing the *correct* answer. Please indicate such references.