

Cognition and Strategy: A Deliberation Experiment*

Eric S. Dickson[†]

Catherine Hafer[‡]

Dimitri Landa[§]

Abstract

A theory of deliberation must provide a plausible account both of individuals' choices to speak or to listen and of how they reinterpret their own views in the aftermath of deliberation. In our experiment, subjects with diverse interests and information choose to speak or to listen and, after updating their beliefs, vote over a common outcome. An important feature of the strategic setting is that not receiving a specific communication is sometimes just as informative as receiving it. Our experimental evidence shows that, although subjects behave instrumentally, their behavior systematically deviates from equilibrium predictions. These deviations indicate a cognitive hierarchy defined by differing abilities to grasp the information inherent in not receiving certain communications, relative to receiving them, and to assess the strategic effects of those implications. We trace the consequences of these underlying cognitive differences for individual deliberative choices and for the informativeness of deliberation.

*We are grateful to Becky Morton, Sanford Gordon, and seminar participants at the NYU Department of Economics for their helpful comments and suggestions.

[†]Corresponding author. Assistant Professor, Department of Politics, New York University. Mailing address: 726 Broadway, 7th floor, Department of Politics, New York University, New York, NY 10003. Email: eric.dickson@nyu.edu. Fax: (212) 995-4184.

[‡]Assistant Professor, Department of Politics, New York University. Mailing address: 726 Broadway, 7th floor, Department of Politics, New York University, New York, NY 10003. Email: catherine.hafer@nyu.edu. Fax: (212) 995-4184.

[§]Assistant Professor, Department of Politics, New York University. Mailing address: 726 Broadway, 7th floor, Department of Politics, New York University, New York, NY 10003. Email: dimitri.landa@nyu.edu. Fax: (212) 995-4184.

1 Introduction

What are the consequences of public deliberation? Should we expect individuals with diverse interests and information to share their information with each other? And if so, how should we expect others to respond to it? Put differently, should we expect deliberation to take place, and when it does, should we expect it to improve the quality of decision-making?

For deliberative democrats, the free exchange of arguments and the responsiveness of individual and collective decision-making to debate have emerged as the new standard of political legitimacy (Cohen 1997). Although the mechanisms supporting this view vary (Elster 1997, Manin 1987, Habermas 1990, and others), most of them share the proposition that the positive effect of deliberation is a consequence of its promoting the quality of individual and/or collective decision-making (Cohen 1997, Fearon 1998). To the extent that post-deliberative decision-making reflects the information or the “best arguments” communicated in the course of interpersonal deliberation, the resulting decisions must be seen as meeting a higher standard of normative acceptability.

The expectation of successful, informative communication that underlies these arguments cannot, however, be taken for granted, and the need to better understand the determinants of such communication is increasingly attracting the attention of game theorists and political economists (Austen-Smith and Feddersen (2005), Gerardi and Yariv (2002), Hafer and Landa (2005, 2006), Meirowitz (2005), Patty (2005), Stasavage (2005), etc.). One of the key questions on the table is an adaptation to the deliberative democratic context of a question that has motivated much of incomplete information game-theoretic work in political science: when do agents find it in their interest to make arguments and to reveal information that is not available to others? To the extent that our interest in asking this question is in explaining and responding to intentional individual behavior, it cannot ultimately be addressed satisfactorily without also asking what the agents do with the information that is available to them. How do they make sense of it? And how do their own expectations of how others make sense of it affect the

behavior of would-be information-transmitters?

The nexus of cognition and strategy identified by these questions and the dynamics of deliberation to which it gives rise is the focus of the experiment we describe in this paper. The experiment is based on a simple strategic framework in which individuals have an opportunity to communicate with one another in advance of a collective decision that is made by simple majority rule. In this framework, individuals may choose either to “speak” or to “listen” as they consider the tradeoffs between influencing others’ choices, on the one hand, and gaining potentially useful information on the other. The players’ capacities to make sense of the information that would, as a consequence of deliberation, be available to them in the voting stage affect their choices of whether to be senders (speakers) or receivers (listeners). That information is of two kinds: messages that credibly, directly, and fully inform the players of the identity of their preferred policy choices, and, when those messages are not received, messages that could, upon reflection (that is, indirectly), be understood as revealing those choices. A famous literary example illustrates the nature of the inference that is entailed in making sense of the latter kind of messages. In Arthur Conan Doyle’s short story “Silver Blaze,” Sherlock Holmes points out that the fact that the dog did not bark in the night implies that the crime was not a burglary. His other claims to genius aside, Holmes’ inference in this and other cases is, as he himself repeatedly announces in Conan Doyle’s stories, straightforward. He says, with a mix of frustration and self-satisfaction at his superiority over the everyman Watson’s failure to keep up, “how often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?”¹

Holmes and Watson differ in their abilities to make an inference from the “null observation.” In the context of deliberation, an equivalent of the “null observation” is the unpersuasive argument, provided that the listener knows something about the correlation between that argument and the position that it is meant to support. Another, more directly politically relevant example helps clarify this point. Suppose that Jill is uncertain in her position on abortion rights: she leans against them, but is not yet convinced that she has heard the most persuasive arguments that would support her position. Suppose, further, that Jill finds herself in a discussion

¹See Conan Doyle (1976a and b).

with someone she has good reason to believe is one of the most thoughtful critics of abortion, and this person makes to her a series of arguments, none of which she finds persuasive. How should this interaction affect Jill's beliefs about the justifiability of abortion rights? If Jill is a Bayesian agent, she should conclude from the conjunction of two facts, her exposure to the most persuasive arguments possible against abortion rights, and the unpersuasiveness to her of those arguments, that abortion rights are more defensible than she had previously thought. If Jill is, however, more like Watson in the Conan Doyle stories, she may be expected to fail to make that inference. If the critic of abortion rights that Jill met believes that Jill's position will change only if she hears (direct) arguments that she finds persuasive, then he loses nothing by making his best case to her. If he, however, believes that Jill, who would otherwise be likely to support anti-abortion politicians, would make inferences from the "null observation" (that is, from the fact that she has not heard a persuasive argument in his speech), he may be reluctant to make his arguments to her. And of course, his ability to conceptualize the latter possibility is only conceivable if he himself is capable of making inferences from the "null," that is, if he himself is more like Holmes than Watson.

As this example suggests, individual deliberative choices should be influenced in important ways both by the nature of individual cognition, and by the interaction between cognition and strategy. The complexity of the nexus between these two critical factors in deliberative contexts and the value of isolating the effects of various elements that may be responsible for its form point to the benefits of a controlled socially interactive experiment such as the one we describe below.

Our experiment has two basic goals. The first is to learn about the behavioral features of collective deliberation: what aggregate behavioral patterns or trends we should both expect to realize and take as a guide in positive and normative theorizing about deliberation. The second, and closely connected goal is to explore the microfoundations of judgment formation in deliberation. Different cognitive-behavioral "ideal types," who approach deliberation in different ways, will perceive different incentives that affect not only the way in which those types of agents interpret messages they receive in the course of deliberation, but also the strategic deliberative choices that are to be made. Because different micro-level ways of approaching deliberation

have different macro-level behavioral consequences, observed patterns of individual choices regarding speaking and post-deliberative voting allow us to shed light on the microfoundations of deliberative practice.

One of our central findings is the strong and stable tendency of the subjects to “over-speak”: to communicate considerably more than the equilibrium prediction for the standard Bayesian agents would suggest. (Indeed, the incidence of *overspeaking* is almost twice as high as the incidence of *underspeaking* compared to the equilibrium predictions for the distinct deliberative roles and situations we analyze.) In so doing, subjects, thus, tend to over-expose themselves to the possibility of their audience moving away from the positions they would have preferred them to maintain, but also, and by the same token, to increase the informational value of deliberation.

The distribution of individual choices that gives rise to this pattern of speaking is, however, far from random. To the extent that being convinced by arguments that unambiguously identify one with a particular ideal point in the policy space is more likely to characterize a more ideologically extreme rather than a more moderate agent (Hafer and Landa 2006), our experimental results suggest that, in the aggregate, subjects who are more extreme with respect to their pre-deliberative positions and arguments tend to speak rather than listen, and more moderate subjects display the opposite behavior.

To account for these findings, we turn to the analysis of cognitive-behavioral ideal types and compare the complete individual behavioral profiles to our predictions for the relevant ideal types of agents. Our evidence suggests that the observed pattern of speaking (and overspeaking) is best explained by the presence of two types of subjects: (1) those who deviate from the Bayesian ideal by systematically displaying a Watsonian approach to belief updating (to identify the right alternatives, they, like Watson and unlike Holmes - the indisputable Bayesian hero - need to see the direct evidence), and (2) those agents who may be best described as “unreflective Bayesians” - though capable of the Bayesian inference themselves, they are unable to appreciate the strategic implications of its possibility.

2 The Model of Deliberation

In this section we describe the strategic framework that we employ in our deliberation experiment. The next section develops theoretical predictions on how fully rational individuals might be expected to behave in the different deliberative situations we employ.² The remainder of the paper then presents and interprets the results of our experimental investigation.

Consider the following sequence of events.³ Individuals who begin in possession of *partial* information about their own best interests are given the opportunity to communicate with one another (the “deliberation stage”); once communication is complete, a vote is held between two potential outcomes, one of which is selected via simple majority rule (the “voting stage”). Each individual receives payoffs that depend on the degree of agreement between her *actual* individual best interests and the election-winning alternative.

Specifically, we consider deliberation within the context of a three-member group. Each member $i \in \{1, 2, 3\}$ has a type $t_i = (t_i^1, t_i^2) \in \{(A, B), (B, C), (C, D)\}$, where $A, B, C, D \in \mathbb{N}$ and $1 \leq A < B < C < D \leq 9$, and an ideal point, or “true number,” $x_i^* = 10t_i^1 + t_i^2$, which corresponds to her most preferred outcome. We use the notation \overline{AB} to denote $10A + B$ - that is, a two-digit number the first digit of which is A , and the second digit of which is B , \overline{BC} and \overline{CD} are defined similarly. The ultimate social outcome is determined by majority rule over a pair of distinct alternatives, $\{y_1, y_2\}$, where $y_1, y_2 \in \{\overline{AB}, \overline{BC}, \overline{CD}\}$. These alternatives are known to all group members from the beginning of the deliberative process. An individual i 's utility from outcome x is linearly decreasing in the “distance” between her true number and the outcome, $u_i(x, x_i^*) = c - |x_i^* - x|$, where c is a constant.

Deliberation has the potential to be persuasive because the players do not know their true numbers for certain at the beginning of the game. Instead, each player initially possesses several pieces of information that are relevant to, but which do not necessarily uniquely deter-

²Our framework is closely related to the model of communication in Hafer and Landa (2005, 2006). Unlike those models, though, the model below incorporates collective decision-making.

³The supplemental appendix provided to reviewers contains the instructions given to subjects, to which an interested reader may refer in order to learn about the way in which this and other aspects of the experimental scenario were presented in the laboratory.

mine, her true number. First, each player knows that the ideal points of every member of her group (including her own) must be drawn from the commonly known set of “true numbers” $\{\overline{AB}, \overline{BC}, \overline{CD}\}$ - for example, $\{13, 37, 79\}$. Second, players know the (unconditional) probabilities corresponding to each of these true numbers. And third, each player knows a “fragment” of her true number, that is, one of the digits from her two-digit true number – for example, “3” or “7” if her true number is 37, but does not know whether that fragment is the first or the second digit of her true number. The set of fragments known to members of the group is common knowledge.

Because, at the outset, a player knows only one of the two digits from her true number, it is convenient to distinguish between the known and unknown fragment. We refer to i 's known fragment as her “active fragment” or “active argument,” $a \in \{t_i^1, t_i^2\}$, and to her unknown fragment as her “latent fragment” or “latent argument,” $l = \{t_i^1, t_i^2\} \setminus \{a\}$.

Individuals' first strategic choice is their mode of deliberative participation $\lambda \in \{0, 1\}$, with $\lambda = 0$ capturing the decision to speak and $\lambda = 1$ the decision to listen. A decision to speak entails an attempt to speak to all other members of her group; a decision to listen entails an attempt to listen to those other members of her group who have chosen to speak. Communication is successful only between individuals who have complementary modes of deliberation – e.g., if i speaks and j listens, communication from i to j takes place. Thus, anyone who chooses to speak receives no messages at all, and anyone who chooses to listen sends no messages at all. When the complementarity of speaking and listening choices occurs, j (the listener) receives a “message” m_j whose nature depends on both (1) speaker i 's active fragment and (2) listener j 's true number. If the speaker's active fragment is part of the listener's true number – corresponding either to the listener's active fragment or to her latent fragment – then the “message” listener j receives is simply speaker i 's active fragment. However, if the speaker's active fragment is not part of the listener's true number, then the “message” the listener receives is that she has received a “foreign fragment” – that is, she has received a fragment that is not a part of her true number (but that does not explicitly indicate what that fragment is). Thus, given $\lambda_i = 0$ and $\lambda_j = 1$, $m_j = a_i$ if and only if $a_i \in \{t_j^1, t_j^2\}$; otherwise $m_j =$ “foreign.”

Before proceeding with the description of the optimal choices in the strategic interaction

described in this section, note, by way of interpretation, that like the players in our game, an individual who participates in political discourse or in other forms of deliberation, can be characterized in terms of the particular set of arguments she may find valid for, or relevant to, the formation of a given judgment. However, similarly to Jill in the example in the introduction, individuals may not always be aware of or have in mind the full set of arguments that they could or would find to be valid or relevant. Communication in which public arguments are expected to convince on the merits - that is, because of the justifiable correspondence of these arguments to accepted facts or premises, *if only of a subset of the audience* - can affect participants' judgments in two ways: (1) by directly bringing to their attention a relevant reason or information that was not an input into the decision-making before, and (2) by indirectly making them aware of the arguments or considerations that they find unconvincing and that, therefore, should not be thought to support their best judgments. Receiving one's latent fragment in our game is, in effect, equivalent to receiving a direct and immediately convincing argument. Receiving a foreign fragment instead of one's latent argument is equivalent to hearing an unconvincing argument. The analysis of a receiver's response to the observation of a foreign fragment, given the information available to her about the nature of the exchange, and of the senders' anticipation of that response yields predictions about the interaction between cognition and strategy illustrated in the introduction.⁴

In our experiment, groups of three subjects each play exactly the game described above, with random re-assignments of subjects and new descriptions of the initial information, consisting of active arguments, true numbers, and probabilities of true numbers, each round. The subjects' communication occurs solely through their terminals, and messages that do not match their

⁴Relative to the existing theoretical literature on sender-receiver games, this strategic framework can be described as one with fully provable and complete messages but with non-common veridicality (truth content) across agents. Apart from Hafer and Landa (2005, 2006), which have these properties, the most closely related work is Lipman and Seppi (1995) who analyze a sender-receiver model in which senders can supply partial proofs for their signals, and Glazer and Rubinstein (2005), in whose model the messages contain "hard" or fully provable information, but this information is partial and the informativeness of messages is a function of speaker credibility. Both of those papers, however, assume common veridicality.

types are transmitted to them as notices of foreign fragments.

From the standpoint of the experimental analysis, one of the key advantages of the deliberative framework described above is that it enables the analyst to rigorously specify a controlled environment for deliberative exchange, including, in particular, the precise factors determining or affecting persuasion. Such a specification is effectively infeasible in natural-language, “real issue,” deliberation. While the analysis of the latter kind of deliberation yields insights about the aggregate consequences of deliberation in a given setting (see, e.g., Fishkin and Luskin 1996; Fishkin, Luskin, and Jowell 2000; French and Laver 2005), our ability to isolate particular determinants of persuasion and trace through their individual consequences is inherently limited by the complex and often idiosyncratic cross-issue linkages and the participants’ responsiveness to a number of more or less apparent cues that go beyond the objectively describable epistemic inputs to the interaction. By removing (or controlling for) those complications, our experiment enables us to explore individual-level behavior in a deliberative environment with a considerable degree of precision. In this respect, the impulse behind our design strategy is similar to that of several experiments reported by Lupia and McCubbins (1998), which study persuasion and choice under a stylized representation of political communication. However, the specific questions we consider and the details of our experimental setup are quite different. For example, we focus on a tradeoff between speaking and listening among members of a deliberative group who must decide whether or not to transmit their partial information about their disparate private interests – and how subjects learn from this interchange – whereas a different focus of Lupia and McCubbins is the distinction between deceptive and non-deceptive communication between an agent and her principal, and the conditions under which such communication will be taken to be credible.

3 Theoretical Predictions

Given the deliberative environment described in the previous section, how can individuals be expected to behave? Because we are interested in the microfoundations of judgment formation in deliberation, it is useful to begin with the baseline standard Bayesian agents as a way of

organizing the data (Schotter 2005). We present the intuitions and the formal behavioral predictions for such agents below.⁵

The following examples illustrate the nature of inference expected from a Bayesian agent in our experiment.

Example 1. Suppose that it is commonly known that the set of true numbers in the group of three players is $\{13, 37, 79\}$ and that the group members' active fragments are 1, 3, and 3. A given player with the active fragment 3 may have either 13 or 37 as her true number, because both of these contain the fragment 3 (but the true number 79 does not). Suppose further that she chooses to listen, and is told that she has received a "foreign fragment." In what way can such a message prove informative for her? The other player with fragment 3 could not have sent the foreign fragment – if that player had sent her fragment, our listener would have been told that she had received the fragment "3" because "3" is part of her true number. Thus, the "foreign fragment" must have been sent by the subject with the fragment 1; hence the foreign fragment must have been 1; hence "1" must not be part of the listener's true number, and thus her true number must be 37. ■

It is clear from this example that, in some circumstances, the receipt of a "foreign fragment" can be as informative as the receipt of the latent fragment. Further, the extent of deduction that is required in order to see this is not particularly demanding. The following two examples illustrate how the possibility of such inferences from the receipt of a "foreign fragment" affects optimal choices regarding speaking vs. listening.

Example 1a. As in the example described above, suppose that it is commonly known that the set of true numbers in the group of three players is $\{13, 37, 79\}$ and that the group members' active fragments are 1, 3, and 3, but also that the ultimate vote for a social outcome will be between 13 and 37. Suppose further that there is common knowledge within the group that 37 is more likely than 13; as such, in the absence of further information, the expected utility-

⁵It is worth emphasizing at the outset the very heuristic nature of these predictions. After presenting the results of our experimental investigation in the subsequent sections, we explicitly consider other cognitive-behavioral ideal types in providing an account of the microfoundations of individual choice that is potentially more consistent with the observations in our experimental data.

maximizing choice for players with the fragment 3 is to vote 37 over 13. Consider first the incentives of the player with fragment 1 in choosing a deliberative strategy. If the players with fragment 3 both speak, the deliberative choice of the fragment 1 player is of no consequence. But that player is weakly better off speaking if at least one of the players with fragment 3 chooses to listen: if the latter receives a message indicating that 1 is her latent fragment, she will know that 13 is her true number, and will now vote for 13 instead of 37 – an improvement for the player with fragment 1 as this now ensures a majority for 13 over 37; if, instead, a listening fragment 3 player receives a message that she has received a foreign fragment, the player with fragment 1 is neither helped nor harmed. That is because a player with fragment 3 will learn that 37, not 13, is her true number *but she would have voted for 37 anyway* in the absence of communication because 37 is more likely than 13. Thus, it is a weakly dominant strategy for a player with fragment 1 to speak in this setting. ■

Example 1b. Suppose now exactly the same setting with one exception: it is commonly known within the group that 13 is a priori more likely than 37. The incentives facing the agents are now different. Sending fragment “1” cannot help, and it may hurt, since a Bayesian who receives a foreign fragment will choose 37. If the sender understands that the recipient may make such a deduction, she will strictly prefer not to send in such a situation.■

We consider two different compositions of deliberative groups (i.e., configurations of known fragments, also referred to below as deliberative situations) in which players face the incentives illustrated in the above examples. These configurations, each corresponding to the set of endowed fragments initially revealed to the players, are *ABB* and *ABC*: for *ABB*, one player with the “A” fragment and two with the “B” fragment, and for *ABC*, one player with “A,” one with “B,” and one with “C” fragments, respectively.⁶ (Note that situation *ABB* and *ABC*

⁶By using three different true numbers instead of two, we can compare the behavior of agents who are and are not certain of their true numbers but who face the same strategic incentives in the group. The diversity of types also allows us to create strategically equivalent situations in which “right” and “left” are reversed, to control for any predilections the subjects may have for choosing or advocating left or right positions, or lower and higher positions. *ABB* (or *CCD*) and *ABC* (or *BCD*) represent the potentially most interesting combinations of active argument types. Because only adjacent types can persuade or be persuaded by each other, in groups that

are symmetric to CCD and BCD , respectively, and so the analysis of the former two situations extends to the latter two symmetrically as well. In what follows, whenever we refer generically to ABC , we mean ABC or BCD , and similarly to ABB meaning ABB or CCD .) We employ notation such as $Pr(\overline{AB}|B)$ to indicate, for example, the (conditional) probability that an agent with active fragment B has true number \overline{AB} .

In both ABB and ABC situations, players vote on \overline{AB} vs. \overline{BC} . Given these voting choices, for all non-degenerate configurations of probability distributions over true number types, players endowed with fragments A and C in the ABC situation and with the fragment A in the ABB situation will always and with certainty know their true number to be one of the corresponding endpoints of the interval $[\overline{AB}, \overline{BC}]$, whereas B agents' *expected* true number will be at a point in the interior of that interval. In a natural sense, then, we can think of the A and C fragments as “extreme” and of the B fragment as “moderate,” with the corresponding reference to the players initially endowed with these fragments as “extremists” and “moderates,” respectively.

Our solution concept is weak dominance, which generates here unique predictions with respect to both speaking vs. listening, and voting. Table 1a below summarizes the equilibrium predictions for speaking and listening faced by the players with the corresponding extreme and moderate fragments in the ABB and the ABC situations under different relative values of $Pr(\overline{BC}|B)$ and $Pr(\overline{AB}|B)$.⁷

TABLE 1a ABOUT HERE

Table 1b below provides our predictions for voting choices given the deliberative situation, the relative values of $Pr(\overline{BC}|B)$ and $Pr(\overline{AB}|B)$, the active fragments, and the possible profiles of messages received given the decision to speak.

TABLE 1b ABOUT HERE

include, for example, a D player but not a C player, or a C player but not a B or D player, at least one player's deliberative behavior must be inconsequential regardless of what other players do. We have excluded such groupings in order to maximize the amount of relevant data that we can obtain in each experimental session.

⁷The formal arguments for the optimality of these strategies are in the Appendix.

As is clear from the discussion of the informal examples in this section and the optimal decisions presented in the tables above, these decisions are characterized by the existence of broad patterns that can be instructively summarized as follows. When, as in Example 1a, (Bayesian) agents would perceive an incentive to speak, we will refer to this as being *Case I* for the relevant *A* or *C* agent (in the *ABB* or *ABC* situation). And when, as in Example 1b, such agents would perceive an incentive to listen (so as not to alienate other favorably disposed agents), we will refer to this as being *Case II* for the relevant *A* or *C* agent (in the *ABB* or *ABC* situation). We will exploit this distinction in our examination of the microfoundations of individual deliberative behavior in Section 6 below.

4 Experimental Sessions

The experiment was carried out at the Center for Experimental Social Science (CESS) laboratory at New York University. Subjects interacted anonymously via networked computers; the experiment was programmed and conducted with the software z-Tree (Fischbacher 1999).

The results for our experiment come from data collected during two experimental sessions involving 18 subjects each, for a total of 36 experimental subjects.⁸ The subjects, who were mostly NYU undergraduates, signed up for the experiment via a web-based recruitment system operated by CESS.⁹ All our subjects gave informed consent according to standard New York University human subjects protocols.

Each of our experimental sessions consisted of 30 rounds, where each round corresponds to a single play of our deliberation game, involving both a deliberative stage and a voting stage. Sessions lasted approximately 90 minutes, and on average subjects earned US\$26.56, including a showup fee of US\$7. At the beginning of each session, a hard copy of the experimental

⁸The data presented here do not include the results of an earlier, pilot session. The pilot session involved a different experimental scenario, involving among other things much more complicated strategy spaces than those described here. As such, we do not consider pilot session outcomes to be meaningfully comparable to our data.

⁹The website <http://experiments.cess.fas.nyu.edu> contains more information about this recruitment system.

instructions (which can be found in the supplemental appendix) was distributed to each of the subjects, and the instructions were also read aloud by an experimental proctor in an attempt to induce common knowledge of the experimental scenario. After the instructions were read, subjects took an on-screen quiz consisting of six questions in order to test their understanding and as a further means of inducing common knowledge of the deliberation framework; subjects were given immediate feedback as to the correct answers to the quiz questions. Once the quiz was complete, 30 rounds of the deliberation game were played. After all 30 rounds were complete, subjects then filled out a debriefing questionnaire.

The 30 rounds of the experiment placed the subjects in a variety of different deliberative situations and assigned them to a variety of different experimental roles. A list of 30 different deliberative situations was compiled, with different sets of true numbers and different unconditional probabilities corresponding to those true numbers. In total, the distribution of situations was as follows: 7 Case I (ABB); 7 Case II (ABB); 8 Case I (ABC); 8 Case II (ABC). In a given round of an experimental session, all of the subjects faced the same situation (although of course they were assigned to different deliberative roles in the context of those situations). In each session the situations were presented to subjects according to a sequence that had been randomly generated by a computer subject to the constraint that each of the first 12 periods (1-12) and the second 12 periods (13-24) contained the following distribution of situations: 3 Case I (ABB); 3 Case II (ABB); 3 Case I (ABC); 3 Case II (ABC). In each of these two blocks of 12 periods, each subject played each role in each situation. This was done in order to ensure that the distribution of the different situations was reasonably dispersed across the experimental session, and to allow for a natural division of each session into a first part (periods 1-12) in which subjects were “inexperienced” and a second part (periods 13-30) in which subjects were “experienced.”

In each period, each subject was rematched into a new group of three subjects, whose actual identities remained anonymous throughout. The matching of subjects into groups, and the distribution of deliberative situations across rounds, was carried out in a sequence that had been randomly generated on a computer, subject to the constraint that each subject was assigned to each deliberative role in each situation an equal number of times both in periods 1-

12 and in periods 13-24. This method of assignment was chosen to ensure that each subject had wide experience of the different deliberative roles over the course of the experimental session.

5 Experimental Results: Aggregate Level

We describe some notable results from the experimental sessions below in the form of a series of conclusions. Each conclusion is followed by a presentation of the data relevant to that conclusion, as well as by some discussion.

5.1 Deliberative Choices: Speaking and Listening

As the discussion in the theoretical results section demonstrates, individuals in our deliberation situation face differing strategic incentives depending on the active fragment that is known to them; the distribution of active fragments across other group members; and the relative likelihoods of different true numbers. Table 2a contains relevant data from our experimental sessions compiled in the same format as Table 1a.

TABLES 2a,b ABOUT HERE

The first thing to note is a systematic difference in deliberative behavior between agents depending on the nature of their active fragment – in particular, whether or not an individual’s active fragment is more “moderate” or more “extreme.” In the settings we describe, listening is always a weakly dominant strategy for individuals with moderate active fragments, while either speaking or listening can be weakly dominant for individuals with extreme active fragments, depending on the situation and their cognitive approach to deliberation.¹⁰ Our first conclusion indicates that the distinction between the deliberative behavior theoretically expected of moderate and extreme agents is strongly apparent in our data.

Conclusion 1. Extremists Speak More, Moderates Listen More. *Subjects with more extreme active fragments speak more frequently than they listen, while subjects with more moderate active fragments listen more frequently than they speak.*

¹⁰Individual decisions are the unit of analysis here as well as in the rest of the paper unless otherwise noted.

The difference in deliberative behavior between “moderates” and “extremists” is striking, as can be seen in Table 2a. Over the last 18 periods of the experiment, once subjects had had some experience of each kind of deliberative role and setting, subjects with a moderate (‘B’) active fragment chose their weakly dominant strategy – to listen – more than 96% of the time in each of the deliberative situations. Pooling across deliberative situations, they did so fully 98.1% of the time (306/312).¹¹ In contrast, subjects with an extreme (‘A’ or ‘C’) active fragment, whose incentives differed across different cases, chose to *speak* between 54.2% and 79.2% of the time in different deliberative situations over the last 18 periods of the experiment. This strong difference in the tendency of moderates and extremists to adopt different modes of deliberation is not only of substantive interest, but also provides one clear piece of evidence that subjects’ deliberative behavior was responsive to the details of their situations.

Of course, the above result for individuals with extreme active fragments combines situations in which such individuals face different tradeoffs between speaking and listening. Our second conclusion reports the way in which their behavior varies between Case I and Case II – that is, between situations in which a Bayesian agent would perceive that speaking as opposed to listening is her weakly dominant strategy.

Conclusion 2. *At the aggregate level, subjects with an extreme fragment chose to speak more often in Case I than they did in Case II. However, they chose to speak substantially more than half the time in both of the Cases, suggesting that population-level average speaking incidence substantially exceeds the predicted incidence for the Bayesian behavioral type. Learning over the course of the experimental sessions significantly increases the incidence of speaking in Case I but does not significantly alter the incidence of speaking in Case II.*

Our most central finding concerning aggregate deliberative behavior can be found, in different forms, in Tables 2a and 2b. Table 2a indicates that, in the last 18 periods, our subjects chose to speak between 54.2% and 66.7% of the time in each of the three settings corresponding to Case II, for an overall average of 60.1% (101/168), in sharp contrast to the Bayesian prediction of 0%. Subjects spoke more often, between 73.3% and 79.2% of the time, in the three

¹¹And four of the six instances in which a moderate spoke during these periods came from a single subject. Note that the corresponding figure over all 30 rounds is 94.9% (501/528).

other settings corresponding to Case I, for an overall average of 76.8% (129/168), as against the predicted 100% for Bayesian agents. The null hypothesis that there is no difference in behavior between Case I and Case II in these last 18 periods can be rejected decisively ($Z = 3.287$, $p < 0.001$). Thus, while predicted deliberative behavior for moderates conforms closely to the theoretical expectations for Bayesians (as seen above, 98.1% observed versus 100% predicted listening), systematic deviations from the Bayesian predictions are observed for extremists, and by far the strongest deviations take the form of “overspeaking” by extremists in Case II.

The importance and robustness of this “overspeaking” in Case II is underscored by considering time trends in subjects’ deliberative behavior. As shown in Table 2b, during the first 12 periods of the experiment, subjects in Case I chose to speak 62.0% of the time (67/108), whereas in Case II they did so 63.0% of the time (68/108) – that is, there is no significant difference in deliberative behavior across Cases when subjects are inexperienced, but as shown above there is a substantial and significant difference once they are experienced. Further, it is important to note that the difference between the first 12 and the last 18 periods is due almost entirely to a shift in subjects’ behavior under Case I. In Case I, the null hypothesis that behavior does not change from the first 12 periods (62.0%) to the last 18 periods (76.8%) is rejected ($Z = 2.636$, $p < 0.01$), whereas in Case II, the null that behavior does not change from the first 12 periods (63.0%) to the last 18 (60.1%) cannot be rejected ($Z = 0.473$, $p = 0.64$). This observation strengthens confidence that subjects’ failure to behave according to Bayesian predictions – and in particular, to exhibit a pronounced pattern of “overspeaking” in Case II – is not entirely a result of inexperienced misunderstanding, because the behavior persists over the course of an experiment with a large number of rounds. That said, it is of course impossible to know how subjects’ behavior would have evolved if it had been possible to run lab sessions with a substantially larger number of periods.

As noted in our theoretical predictions, we employ deliberative settings with different distributions of active fragments; this element of variation in our design gives us further ability to probe the extent to which subjects’ *strategic* incentives influence their behavior, as opposed to other factors that are strategically irrelevant in our framework. A striking feature of our experimental results is the degree of instrumentalism that subjects’ choices may be seen to

reveal, despite their apparent suboptimality. Indeed, we have the following conclusion:

Conclusion 3. Subjects With Identical Strategic Incentives Deliberate Similarly Across Informationally Distinct Deliberative Contexts. *‘B’ agents, who have a dominant strategy to listen, do so at statistically identical rates in ABB and ABC situations. Further, ‘A’ agents and ‘C’ agents who face strategically identical problems within the ABC situation also differ little in their deliberative behavior, even though ‘A’ agents can be ex ante certain of their true numbers while ‘C’ agents cannot.*

Thus, subjects’ strategic incentives matter more to their behavior than do such descriptive aspects of the deliberative problem as the degree of *ex ante* certainty they have about their true numbers. Even when listening is crucial to the subjects’ ability to learn arguments necessary for determining their ideal policies (their unconstrained optima), their decisions regarding speaking and listening tend to be induced by the particular voting agendas - that is, by the extent to which what they could learn in deliberation is *necessary* for determining their “induced” preferences over the alternatives on the agenda (their constrained optima).

The *ABB* and *ABC* situations differ in their degrees of descriptive and inferential complexity – a ‘B’ agent may have more difficulty discerning the meaning of a foreign fragment under *ABC* than under *ABB*, for example, while a ‘C’ agent who has a strategic incentive to send may find this more difficult to understand than a comparable ‘A’ agent, because the ‘C’ agent does not know her true number for certain while the ‘A’ agent does. Table 2a indicates that for experienced subjects in the last 18 periods that most of the variation in deliberative behavior across circumstances is captured by the strategic logic of the model rather than by contextual factors such as these that are strategically irrelevant. ‘B’ agents listen with remarkable consistency across deliberative settings – the variation between 96.7% and 99.0% is statistically insignificant. Similarly, ‘A’ and ‘C’ agents who, as Bayesian agents, would perceive an incentive to speak, also are remarkably consistent across contexts, doing so between 73.3% and 79.2% of the time, another insignificant difference. Finally, ‘A’ and ‘C’ agents who, as Bayesian agents, would perceive an incentive to listen, appear to do so a bit less consistently across contexts, varying from 54.2% to 66.7% of the time. In particular, such ‘C’ agents under *ABC* speak a bit less than do such ‘A’ agents under *ABC*, which may reflect ‘C’ agents’ relative lack of

true knowledge about their actual true number. However, even this difference is statistically insignificant.

In order to further probe the extent to which behavior may have been affected by strategically superfluous descriptive features of the deliberative environment, we carried out a probit regression analysis of deliberative choice for those subjects with an extreme active fragment. Results from the regression, which was carried out for periods 13-30, are contained in Table 3. The dependent variable is the dichotomous choice between listening (1) and speaking (0). The independent variables depict a variety of features of the deliberative environment as well as a time trend variable.¹² The regression results indicate that only subjects' strategic incentives – as categorized by the distinction between Case I and II – significantly affect behavior. Other factors, such as the distinction between *ABB* and *ABC* situations; the distinction between “left-” and “right-handed” situations (e.g., *ABB* vs. *CCD*); and strategically irrelevant features of the true numbers' and probabilities' specific values all have a statistically insignificant effect on behavior. The time trend variable is also insignificant, indicating no strong learning trend after period 13. These results give us confidence that subjects' choices were affected by their strategic incentives but were not unduly influenced by other factors. With these results in mind, for the remainder of the section we separately pool all decisions together that are within a common case (I or II), and restrict our attention to periods 13-30.

TABLE 3 ABOUT HERE

5.2 Voting Choices

We begin our consideration of the voting data by noting the frequencies with which subjects who play a particular role in a given round of the experiment are exposed to a variety of different deliberative outcomes. These frequencies are of course determined jointly by chance and by the deliberative strategies chosen by a given subject and his or her counterparts. For a given individual who has been assigned the ‘*B*’-type fragment, three separate results of deliberation are about equally likely to be observed: the receipt of no fragment (29.9% of the time); the

¹²Definitions of the independent variables can be found in the text accompanying Table 3.

receipt of a foreign fragment only (26.7% of the time); and the receipt of the subject’s latent argument only (25.0% of the time). The conjunction of the latent argument and a foreign fragment was considerably less frequent (15.7% of the time), and the rest of the possible outcomes essentially not present (from at or below 1% of the time).

The frequencies of the deliberative outcomes perceived by individuals who had been assigned the ‘A’- or ‘C’-type fragments reflect the different deliberative choices made by these agents compared to those who possess the ‘B’ fragment. Most of the time (84.2%) ‘A’ or ‘C’ agents receive no signal at all – naturally the case as we have seen that ‘B’ agents almost always listen and the ‘A’ and ‘C’ agents themselves speak more often than they listen. ‘A’ and ‘C’ agents receive only a foreign fragment 13.6% of the time and observed other outcomes very rarely.

Along with our theoretical expectations, these distributions of deliberative outcomes lead us to expect we will learn most about subjects’ behavioral types from voting behavior by observing the choices made by the ‘B’ agents. Our next two conclusions focus on those choices.

Conclusion 4. Subjects Almost Always Use Dominant Voting Strategies When Receiving No Signal or When Receiving Their Latent Fragment. *At the aggregate level, when subjects with a ‘B’ (moderate) active fragment receive no signal, they almost always vote according to their efficient prior belief, indicating a strong understanding of the probabilities’ actual meanings. Further, when subjects with a ‘B’ active fragment receive their latent fragment, they almost always vote correctly (for their actual true number).*

TABLES 5a,b ABOUT HERE

The data supporting this Conclusion can be found in Table 4a. Aggregating across deliberative settings, subjects with a ‘B’ (moderate) active fragment who receive no signal at all vote in accordance with their prior belief 94.9% of the time (150/158). Further, subjects with a ‘B’ active fragment who receive their latent fragment vote correctly (for their now-known true number) 95.0% of the time (209/220). Taken together, these statistics indicate a striking degree of understanding among our subjects of the meaning of the unconditional probabilities they were given and of the meaning of latent fragments.

With this information in hand as a baseline, the following conclusion describes voting be-

havior by those ‘*B*’ agents who receive only a foreign fragment.

Conclusion 5. Subjects Often Fail to Learn From Informative Foreign Fragments. *At the aggregate level, when subjects with a ‘*B*’ active fragment receive only a foreign fragment in Case II (*ABB*), the pattern of voting behavior indicates a strong and systematic failure to infer that their true numbers are not \overline{AB} . In the cases in which receiving a foreign fragment should indicate to subjects that they should vote according to their priors, they vote correctly almost all of the time. But in the cases in which receiving a foreign fragment should indicate to subjects that they should vote against their priors, they vote correctly much less often (slightly more than half of the time).*

This conclusion exploits the differences in the information content of foreign fragments in different deliberative settings. In Case I (*ABB*), the ‘*B*’ agent’s prior indicates that, in the absence of further information, she should prefer to vote for \overline{BC} over \overline{AB} . If, in this setting, a Bayesian ‘*B*’ agent receives a foreign fragment, she will understand that the foreign fragment must have been sent by the ‘*A*’ agent; that her own true number cannot contain the ‘*A*’ fragment; and therefore that her own true number must be \overline{BC} . The same is also true in an *ABC* situation: a Bayesian agent who understands the strategic incentives faced by ‘*A*’ and ‘*C*’ types will expect that, if one foreign fragment only is received, that it would have most likely been sent *by the agent who was trying to change her mind* – and therefore should not affect her ultimate vote.

However, the situation in Case II (*ABB*) is quite different. The ‘*B*’ agent’s prior belief inclines her to vote for \overline{AB} over \overline{BC} . However, unlike in the cases discussed above, receiving a foreign fragment puts the ‘*B*’ agent’s prior beliefs in tension with the information represented by it. She now ought to vote against her prior belief, because she would know with certainty that her true number could not contain fragment ‘*A*.’ As such, a comparison of the voting behavior of ‘*B*’ agents who have received only a foreign fragment between Case II (*ABB*) and the other cases provides a direct test for the hypothesis that agents fail to make inferences from unpersuasive but informative arguments. The results of this comparison can be found in Tables 4a and 4b.

The rate with which subjects fail to vote “correctly” upon receiving a foreign fragment

that indicates against their prior is striking. Taken over all 30 periods, subjects vote correctly (against their prior) 56.2% of the time (18/32) in Case II (*ABB*), but they vote correctly (with their prior) 91.7% of the time (100/109) in the other cases. The null hypothesis that these success rates are identical is soundly rejected by a difference-of-proportions test ($Z = 4.778$, $p < 0.0001$). Restricting attention to the last 18 periods (13-30), as can be seen in Table 4b, subjects vote correctly (against their prior) 62.5% of the time (10/16) in Case II (*ABB*), but they vote correctly (with their prior) 93.4% of the time (71/76) in the other cases. In this instance as well the null hypothesis that these success rates are identical is overwhelmingly rejected ($Z = 3.465$, $p < 0.001$).

At the risk of repeating ourselves, it is worth emphasizing that the high rate of errors in voting that we note above takes place in the context of Case II (*ABB*). If a subject receives a foreign fragment it is *only* the ‘A’ agent from whom the foreign fragment could have come. While the information contained in such a foreign fragment is not as transparently *labelled* as the other sorts of information – the receipt of a latent fragment, or being endowed with an “endpoint” fragment – that lead to correct voting rates in excess of 95%, it is in actuality no less informative. The ability of subjects to make correct inferences where they do, but *not* to make them when receiving such foreign fragments, fits very well with the logic of a violation of *Negative Introspection* (the ability to learn from what does *not* happen or what one does *not* know).¹³

6 Re-examining Microfoundations

6.1 Cognitive Types and Individual Behavioral Profiles

The five conclusions presented above all concern aggregate-level data that pools together the behavior of all of our experimental subjects. Of course, aggregate-level findings about deliberative choices may result from a wide variety of different individual-level behaviors. Among other things, the given aggregate-level findings can be generated either by a homogeneous population of individuals who behave probabilistically in the same way that population averages do, or by

¹³See, e.g., Binmore 1990, p. 108-110.

various configurations of highly heterogeneous individual behavior. In this section we address our experimental findings at the individual level. Although the conclusions reported in the previous section indicate substantial presence of suboptimal behavior, they also show that the behavior is far from random, naturally raising the question of what may account for it.

As we noted in the introduction, our explanation hinges on individuals' understandings of the epistemic and strategic implications of indirect evidence, variation in which gives rise to distinct cognitive types in the population of subjects. Before examining the individual-level data in detail, we first consider and reject another account that may, at first blush, seem consistent with the evidence presented in the previous section. This account turns on the presence of a population of agents who have so little understanding of the problem that the unconditional probabilities of the true numbers have no meaning for them at all. Such agents might not vote according to the relative values of the unconditional probabilities – that is, they may ignore the distinction between Cases I and II, perhaps voting merely randomly. However, even such agents could probably be expected to understand their actual true numbers if they possessed both the active and the latent fragments. If present in the population in sufficient numbers, such agents might induce Bayesian subjects to behave differently in equilibrium than our theoretical expectations indicate. In particular, a Bayesian agent might have an incentive to speak – regardless of the Case, and so potentially consistent with our finding of considerable “over-speaking” in Case II – if she believed that the population of such deviant agents was so large as to make the advantages of persuading them outweigh the potential risks of dissuading fellow Bayesians.

There are several reasons that point to the implausibility of this explanation. First, subjects virtually universally gave correct answers to the question regarding the unconditional probabilities on our quiz (34/36, 94.4%).¹⁴ Second, as noted above in Conclusion 4, subjects with the

¹⁴The relevant quiz question reads as follows: “Suppose that the set of possible true numbers is 26,68,89. Suppose that the frequency of the true number 26 is 35%, that the frequency of the true number 68 is 45%; and that the frequency of the true number 89 is 20%. And suppose that a person in your group is told that 6 is a fragment of his or her true number. What is his or her most likely true number given this information?” The feedback involved reminding subjects of their own answer, and then adding: “The correct answer was 68. If the person has

moderate fragment who receive no signal at all vote according to their (unconditional) prior belief practically all of the time. Taken together, these two points indicate that very few agents actually possess the particular deficiency in understanding described above that would give Bayesian agents the incentive always to speak. Beyond that, subjects' responses to the questions in our post-experiment debriefing questionnaire provide further skepticism regarding the plausibility of this explanation. (We discuss several aspects of these responses, including their high correlation to the behavior observed in the experiment, below.) When asked "Did you find the problem at hand difficult or easy?" and "Do you think other people found the problem difficult or easy?" only 4 of the 36 subjects (11.1%) give responses suggestive of thinking that they found it easier than their counterparts, and only one of these four subjects exhibits deliberative and voting behavior that is consistent with being a Bayesian agent with the belief about her counterparts that is consistent with this explanation. This pattern of responses stands in sharp contrast to what one would expect if Bayesian subjects believed that much of the population consisted of rogue deviant types who could not properly understand the deliberative environment, and who must therefore be sent fragments as often as possible in hopes of activating their latent fragment.

The implausibility of the explanation that relies on agents' beliefs that their counterparts are "too irrational" points to the appeal of an explanatory strategy we pursue in the remainder of this section. This strategy is to identify analytically plausible cognitive-behavioral ideal types and compare our predictions for their respective behaviors with the complete individual behavioral profiles of our experimental subjects. A natural point of departure in constructing such ideal types is the Bayesian agents that appear in standard game-theoretic models and that underlie our initial theoretical predictions. Such agents, who update their beliefs efficiently using Bayes' Rule and who, like Holmes, understand the logical implications of any information they possess and incorporate them into their beliefs, may be thought of as being at the top step of the hierarchy of different types that possess different levels of understanding of the strategic environment and use the information they receive with different levels of insight.

a fragment 6, his or her true number must be either 26 or 68 because 89 does not contain the fragment 6. And 68 occurs with greater frequency than 26."

Agents below the top step in this hierarchy differ from the Bayesian agents in that they lack some degree of logical omniscience: they are unable to recognize some of the logical implications of certain pieces of information they encounter, and, because of this, make suboptimal choices. In the context of the analytical tasks faced by our subjects, there is a natural line that can be drawn between those agents who are capable of making a Bayesian inference from a foreign fragment and those who are not. With this line in mind, we can define a cognitive type - we refer to it as *Watsonian* - that differs from the Bayesian ideal type in that agents that belong to it are not negatively introspective, i.e., they do not make the necessary inferences in order to correctly update their beliefs upon receiving a foreign fragment, even though (as we have seen) foreign fragments can be informative. Like Holmes' Dr. Watson, these agents are in other ways quite smart: they understand all other aspects of the deliberative environment, aside from a knowledge of how they (or others) could use the information contained in foreign fragments.

Various strands of the cognitive and social psychology literatures suggest that the expectation of the presence of Watsonian agents is, indeed, an empirically reasonable one. A wide variety of experimental studies indicate a cognitive bias in favor of one's own currently held convictions; one consequence of such a bias can be that only direct and largely unambiguous evidence is able to overturn one's prior position (Zaller 1992; Dawes 1998; Rabin 1998; Baron 1994). As the authors of one of the seminal studies note, agents "may even come to regard the ambiguities and conceptual flaws in the data opposing their hypotheses as somehow suggestive of the fundamental correctness of those hypotheses" (Lord et al. 1979, pp. 2099). Perhaps most relevantly, the literature on the psychology of hypothesis testing (Wason 1968, 1977; Baron 1994, Ch. 13) suggests that people tend to look for "positive" confirmations of hypothesized patterns while disregarding or failing to look for "negative" signals which do not fit the expected pattern but which can disconfirm the hypothesis. In our framework, such a phenomenon would be observed if a Watsonian with a given prior belief about her most likely true number were to fail to update away from this belief upon receiving a "foreign fragment" that logically falsifies it. Still, the prior experiments that would support this expectation are substantially different from ours in their cognitive demands on the subjects.¹⁵ They are also,

¹⁵Thus, for example, the various versions of the seminal "card experiment" that are the

and perhaps most importantly, set in different decision-making environments. To the extent that individual inferences may be affected by, inter alia, the observations of and beliefs about the choices by one’s strategic counterparts, whether the expectations induced by the behavior in the decision-theoretic settings are sustained in interactive settings is not *ex ante* clear.

Indeed, the very interactive nature of the decision setting in our experiment suggests a possibility of a further cognitive type that falls between Watsonians and Bayesians in the cognitive hierarchy. To see the intuition for it, note that the analytical tasks facing our subjects may be seen as nested in the following sense: the correct identification of the optimal speaking behavior for the Bayesian agents effectively presupposes the correct solution of a somewhat easier problem of inference as a listener. The difference in difficulty between these problems suggest the possibility of a cognitive-behavioral ideal type that is capable of a negatively introspective inference herself, but different from the standard Bayesian type in being unable to make the inferences from that possibility for the strategic choice of speaking vs. listening. We refer to the agents who belong to this ideal type as *Unreflective Bayesians*.

Bayesians, Unreflective Bayesians, and Watsonians do not, course, exhaust the set of possible ideal types. We can also imagine agents who are below Watsonians in the cognitive hierarchy: for example, agents who behave randomly, or who have an idiosyncratic and more gravely mistaken understanding of the deliberative environment than Watsonian agents do. We do not, however, possess strong intuitions for which of the many such possible idiosyncratic behaviors are most likely to be observed in actual subject populations, and so do not provide a specific characterization of the behavior we would expect of such agents, referring to them instead collectively under the single label of *Deviant*.

Because neither Watsonian nor Unreflective Bayesian agents are able to conceive of the possibility of their fellow group members using “against them” the information they might provide

core of the experimental hypothesis-testing literature require subjects to evaluate the truth of conditional statements and determine what pieces of information they wish to acquire given what can be inferred from every conceivable piece of information. By contrast, the immediate inferential task facing the subjects who choose to listen in our experiment is simpler: they must evaluate the truth of an unconditional statement (e.g., “my ideal point is \overline{AB} ”) and are confronted with a specific piece of evidence (e.g., a foreign fragment).

by choosing to send, they are indifferent between sending and receiving in all circumstances in which they are initially endowed with an extreme fragment and the moderate agent’s pre-communicative beliefs are in their favor. In Table 1a, which describes the equilibrium speaking and listening choices of Bayesian agents, these circumstances correspond to the choices marked in boldface. In fact, these choices are the only differences in optimal speaking sub-strategies between them and the Bayesian agents. (The formal derivation of the Watsonians’ optimal strategies is in the Appendix; the Unreflective Bayesians’ speaking sub-strategies are the same and are justified by the same argument). The only difference between Watsonians and Unreflective Bayesians concerns the voting choice in Case II (ABB) when receiving a foreign fragment - i.e., the situation that corresponds to the bold font entry in the Table 1b. Whereas the Bayesians and the Unreflective Bayesians would switch from the prior optimum of \overline{AB} to \overline{BC} , Watsonians would continue preferring \overline{AB} .

Given these predictions, we can now look for evidence for each of these ideal types in the individual-level data. Keeping in mind the observational equivalence between the Unreflective Bayesians’ and Watsonians’ deliberative choices (i.e., speaking vs. listening decisions), we can state the following conclusion:

Conclusion 6. Inferences about Behavioral Types: Individual Deliberation Data.

At the individual level, there is substantial heterogeneity in the deliberative behaviors demonstrated by subjects when cast in the role of extremist. Roughly half of the subjects demonstrate behavior that is consistent or nearly consistent with the behavior theoretically associated with the Watsonian and the Unreflective Bayesian behavioral types. A much smaller number of subjects demonstrate behavior that is consistent or nearly consistent with the behavior theoretically associated with the Bayesian behavioral type. The remaining subjects display highly idiosyncratic behaviors, some of which involve the repeated selection of dominated strategies.

TABLE 5 ABOUT HERE

Table 5 contains the individual-level deliberation choices in rounds 13-30 for subjects who possess extreme active fragments. (Note: subjects 1-18 took part in Session 1; subjects 19-36 took part in Session 2.) We restrict our attention to rounds 13-30 in an attempt to obtain

a relatively meaningful, if rough, classification of behavior that takes into account an initial period of learning by subjects.¹⁶ In order to organize this data, we employ the following system of categorization. If a subject behaves in precisely the same way as a given behavioral type would, or if the subject deviates from this behavior by no more than one choice in Case I and by no more than one choice in Case II, the subject is described as being of that behavioral type. For the Bayesian type, this involves speaking in all or all but one of the Case I circumstances, and listening in all or all but one of the Case II circumstances. For the Watsonian and the Unreflective Bayesian type, this involves speaking in all or all but one of the Case I circumstances, and speaking in at least half of the Case II circumstances.¹⁷ Behavior that is not classified according to either standard will be discussed later.

This system of categorization yields the following results. Of the 36 subjects, 19 can be uniquely categorized as belonging to the Watsonian/Unreflective Bayesian types. Of these 19 subjects, 15 make choices that correspond exactly to possible choice behavior for a Watsonian and the Unreflective Bayesian agents as we have defined them; 11 subjects speak in every round, whether it be an instance of Case I or of Case II, while 4 subjects speak in every round of Case I and in at least half of, but not all of, the instances of Case II. The remaining 4 subjects deviate from one of these two patterns by one choice according to the definition prescribed above.

At the same time, only 3 subjects can be uniquely categorized as belonging to the Bayesian type; all 3 of these subjects speak in every Case I situation but listen in every Case II situation. These unambiguous assignments of subjects to the Watsonian/Unreflective Bayesian or to the Bayesian type account for 22 of the 36 subjects. Of the remaining 14 subjects, 2 subjects “fall on the boundary” between Bayesian and Watsonian/Unreflective Bayesian types; because these subjects were exposed to a small number of instances of Case II, they were both simultaneously one choice away from the Bayesian classification and one choice away from the Watsonian/Unreflective Bayesian classification. As such, these subjects are left uncategorized.

¹⁶Although the specific classifications, as well as the qualitative character of the following discussion, would have differed rather little even if we had used the data from all 30 periods.

¹⁷Note that this is a conservative standard that cuts against the affirmative classifications of the Watsonian and the Unreflective Bayesian types.

Two other subjects fell one further choice away from one of the behavioral types than the categorization system allowed; one of these was “leaning” towards a Bayesian classification and the other was “leaning” towards a Watsonian/Unreflective Bayesian classification. The remaining 10 subjects did not fit the pattern of either behavioral type, engaging in especially perverse behavior – for example, always listening, regardless of the case – or exhibiting a pattern of choices that appears essentially random. In Table 5 these subjects are labelled “deviant”; the table contains the categorizations of all of the other subjects as well. Figure 1 provides a graphical representation of the data on deliberative choices; the radius of each circle in the figure is proportional to the number of subjects at a given point in the space.

Based on the deliberation decisions that our subjects make, Conclusions 2 and 6 provide clear aggregate- and individual-level indications that a view of our subjects as a homogeneous population of Bayesian types cannot be supported. Although a small number of subjects do appear to choose whether to speak or to listen in a way that more or less corresponds to what we would expect of Bayesian agents (in our data, 3 or 4 out of 36), much the larger number demonstrate tendencies to choose deliberative strategies that a true Bayesian type would perceive to be dominated by other alternatives. While the subject population appears to be heterogeneous in the way its members approach the deliberative problem, the largest number of subjects behave in a way that corresponds according to our classification system to the Watsonian/Unreflective Bayesian types we postulate (in our data, 19 out of 36).

We next consider how the hierarchy of ideal types we proposed above comports with subjects’ voting behavior. Because virtually all subjects vote consistently with their prior in the absence of a message informing them of having received a foreign fragment, the relevant variation concerns the circumstances in which they do receive such a message. Table 6 reports individual-level information regarding subjects’ voting choices. Before considering it, note that obtaining the necessary sample size to make robust statistical inferences is difficult for several reasons. In order to receive a foreign fragment in Case II (ABB), it must be the case that a given moderate agent chooses to listen; that the extreme agent chooses to speak; and that the extreme agent’s active argument is not part of the moderate agent’s true number (that is, the moderate agent’s true number is \overline{BC}). But in Case II (ABB), the moderate agent’s true number is more likely

actually to be \overline{AB} than it is to be \overline{BC} – so instances in which the moderate agent receives a foreign fragment rather than the fragment A are comparatively infrequent. And, of course, at least from the Bayesian perspective, it is out-of-equilibrium for the extreme agent to choose to speak in such a setting in the first place.

Keeping that in mind, our analysis of the individual-level voting data gives rise to the following conclusion:

Conclusion 7. Consistency of Behavioral Types Between Deliberation and Voting. *At the individual level, when correlating deliberative behavior with voting behavior, subjects who behave as Bayesian agents in choosing a deliberative strategy vote correctly upon receiving only a foreign fragment. When pooled together, subjects who behave as Watsonian/Unreflective Bayesian agents in choosing a deliberative strategy vote correctly less often but more often than the subjects classified as Deviant.*

If the behavioral types we have defined are a reasonably good description of how subjects view deliberation and form judgments about their interests, and they correctly instantiate a true latent cognitive-behavioral hierarchy present in the population of subjects, it should be the case that the deliberative choices made in the first stage exhibit consistency with the voting choices made in the second stage. In particular, subjects who behave as Bayesians would in the deliberative phase – responding to the differing incentives induced by a fear of driving away likely supporters on the one hand but swaying likely opponents on the other – should also demonstrate a correct understanding of the meaning of foreign fragments in the voting phase. Agents who were classified as Deviant in the deliberative phase should fail to learn as much as they should from foreign fragments when they cast ballots, and the degree of correctness in post-deliberative choices by the subjects who behaved as Watsonian or Unreflective Bayesian agents in the deliberative phase should, when pooled, fall between the values for the Bayesian and the Deviant agents. (The more Unreflective Bayesians are in that group, the closer the group value should be to that for the Bayesians, the more true Watsonians there are, the further it should be.)

In order to test for this linkage, we compared subjects’ individual-level vote choices for the 32 instances of Case II (ABB) contained in Tables 6a and 6b with their behavioral classifications

– derived from deliberative choices made in different rounds – contained in Table 4. The result was as follows: agents classified as Bayesian voted correctly 100% of the time (4/4); agents classified as Watsonian and Unreflective Bayesian voted correctly 57.9% of the time (11/19); and agents classified as Deviant voted correctly 28.6% of the time (2/7).¹⁸ As such, the pattern in subjects’ ability to use information from foreign fragments in order to vote correctly seems to be correlated with the behavioral classifications from deliberative choice in the way one would expect if stable behavioral types were an accurate model of subjects. However, given the small sample size of situations in which ‘*B*’ agents received only a foreign fragment in Case II (*ABB*), it is not possible to make a strong statistical claim that this correlation did not arise due to chance.

6.2 Learning and Strategic Inference

The data concerning linkages between deliberative behavior and voting behavior, along with observed response patterns in our questionnaire, lead us in the direction of a final conclusion.

Conclusion 8. Significant Presence of both Watsonian and Unreflective Bayesian Types.

The pattern of voting behavior exhibited by subjects who were classified as either Watsonian or Unreflective Bayesian based on their deliberative choices is suggestive. While it is not reasonable, at the individual level, to classify subjects as Watsonian or as Unreflective Bayesian based on their voting behavior (due to the low frequency with which a given individual subject confronts conditions relevant to distinguishing the correct classification), observed voting patterns at the aggregative level lend themselves to an interpretation that Watsonian and Unreflective Bayesian types are likely both present in the population in significant numbers. Indeed, the fact that the relevant pool of subjects votes correctly 57.9% (11/19) of the time can be interpreted as evidence for this; if that pool were homogenously Watsonian this figure would be 0%, while if it were homogeneously Unreflective Bayesian, the figure would be 100%.

That the actual figure is significantly lower than 100% – and significantly lower than the corresponding figure for Bayesian-classified subjects – is compelling evidence for a substantial

¹⁸In addition, “unclassified” agents voted correctly once and incorrectly once.

number of true Watsonians in the population. Given that, as we have seen, the voting behavior of subjects as a whole is nearly always correct outside of the Case II (*ABB*) situation, the presence of a single, sharp downward deviation in voting accuracy in exactly the place it would be expected of Watsonians does seem to make for a strong case that Watsonians are actually present. However, the fact that the figure of 57.9% also significantly deviates from 0% points also to the presence of Unreflective Bayesians in the Watsonian/Unreflective Bayesian camp.

Subjects' responses to the questions in our post-experiment debriefing questionnaire provide further strong support both for this conclusion and for the cognitive hierarchy explanation of the experimental findings more broadly. Our first two questions prompted subjects to describe their approach to making deliberation decisions, while the third question addressed the way in which communication outcomes influenced vote choice.¹⁹ We coded each subject's responses to these questions, blind to actual behavior in the experiment, as exhibiting Bayesian, Watsonian, or Deviant tendencies. 19 of the 36 subjects (52.8%) gave "Watsonian answers" to all three questions; 5 subjects (13.9%) gave consistently "Bayesian answers"; and 3 subjects (8.3%) gave robustly deviant answers. Of the remaining subjects, 5 (13.9%) gave "Watsonian answers" to the deliberation questions but a "Bayesian answer" to the voting question (a pattern consistent with the Unreflective Bayesian type), and the last 4 (11.1%) exhibited other patterns of responses of mixed Bayesian and Watsonian character. Our blind coding of these questionnaire responses was very highly correlated with actual behavior in the experiment, providing further support for the interpretation of subjects' behavior that we explicate above. Of the 4 subjects classified as Bayesian (or Leaning Bayesian) in Table 6, all 4 gave consistently Bayesian answers to the questionnaire questions; of the 20 subjects classified as Watsonian/Unreflective Bayesian (or Leaning Watsonian), 14 gave consistently Watsonian answers, 4 gave Watsonian answers to the deliberation questions but Bayesian answers to the voting question, and 2 gave answers of mixed character.²⁰ The high level of correspondence between our behavioral classifications

¹⁹The text of these questions was as follows: 1. How did you decide when to send vs. receive? 2. Did your choice of sending vs. receiving vary depending on circumstances? If so, how? 3. Did communication often help you decide how to vote - and if so, how?

²⁰In addition, of the 2 unclassified subjects, 1 gave consistently Bayesian responses while the other gave responses of mixed Bayesian and Watsonian character.

on the one hand, and subjects' own depictions of their decision making processes on the other, provides us with substantial direct support for our interpretation of behavior. At the same time, there was little or no evidence in questionnaire responses of subjects' hypothesizing play against deviant types, and choosing optimal strategies given such a hypothesis.

Finally, analysis of the questionnaire responses offered by subjects classified as deviant suggest that our behavioral classifications may somewhat overestimate the proportion of true deviants in the population – and that, in particular, many of these deviants may possess unmeasured Watsonian inclinations. Of the 10 subjects classified as deviant in Table 4, only 3 of them gave robustly deviant questionnaire responses; 5 of them provided responses that were classified as consistently Watsonian; and the other 2 fell into one of the mixed Watsonian-Bayesian categories. The reasons for the inconsistency between observed behavior and survey responses among subjects classified as deviant are unclear. One possibility is that subjects, upon filling out the survey questions, simply grasped for *ex post* rationalizations of behavior that may have had little if anything to do with the actual processes generating the behavior. Another possibility is that deviant subjects may have “learned” either Bayesian or Watsonian behavior, but that they did so sufficiently late in the experimental session that this “learning” could not be behaviorally measured.²¹ In either case, it is noteworthy that the behavioral rationalization – or the behavior towards which learning guided subjects – was much more commonly of Watsonian rather than of Bayesian character.

²¹Though note that the regression contained in Table 3 contains no evidence of aggregate learning in Periods 13-30. Further, numerous additional regression specifications were employed in an attempt to isolate the most likely points at which experience-based learning might have occurred. For example, a subject in the moderate role who chose to listen but received no message at all might wonder whether there could be a strategic advantage to listening as an extremists – and hence stumble onto Bayesian reasoning. However, regressions taking note of subjects' exposure to such potentially enlightening experiences, either from one period to the next or lagged over a varying window of prior exposure, all returned insignificant coefficients on the experience-based learning variable.

7 Conclusion

Constructing empirically relevant theories of deliberation requires that we take seriously how the content of and response to the voiced arguments may be affected by individuals' perceptions of each other in the context of the underlying political issue or problem. What may be expected from this relationship is a function of the particular features of what we described in the introduction as the “nexus of cognition and strategy.” The experiment described in this paper is an attempt to shed light on those features. It offers an account of deliberation that suggests certain aspects of consistency with but also, and perhaps more interestingly, key fundamental and systematic differences from what one might expect on the basis of equilibrium predictions for the underlying strategic model. Although we find that subjects are quite instrumental in their choices in that they pay attention to the payoff-relevant and ignore the payoff-irrelevant features of their choice situations, we also find that they exhibit a persistent tendency to ignore the strategic implications of their speech in making choices about their deliberative participation.

By far the strongest manifestation of this tendency is subjects' engagement in public argumentation when their optimal choices counsel silence. Two consequences of this outcome for deliberative practice deserve particular mention. The first of these consequences is behavioral, and it receives direct strong support in the experiment: the more ideologically extreme agents should, *as a rule*, be expected to speak rather than listen, while the more ideologically moderate agents exhibit the opposite behavior. The second consequence concerns the informational content of deliberation. While “overspeaking” connotes suboptimality of individual deliberative choices, it also indicates that deliberative engagement may lead to greater levels of socially available information for post-deliberative decision-making than would exist under the optimal speaking and listening choices. Individual suboptimality here may have a social silver lining. Still, our results also show that while some agents may be expected to take advantage of this “unexpected” informational windfall, others – in our experiment, almost a half of the participants – may not.

When combined with the expectation of greater speaking by those with the strongest prior

ideological bias, this kind of inefficient learning has an important implication for understanding the conditions that give rise to the phenomenon of post-deliberative group polarization (Mendelberg 2002, Sunstein 2002). The Watsonian failure to learn from “the null event” means that the response to biased speech in closed biased groups is likely to be asymmetric: movement toward the biased speaker is more likely than movement away from her.

Our analysis of individual behavioral profiles provides strong evidence that the nature of the responsiveness and unresponsiveness to “null events” is, indeed, the critical systematic factor that explains the variation in subject behavior. But while a considerable subset of subjects displays the quintessentially Watsonian epistemic unresponsiveness to such events, we also find evidence of a fundamentally strategic type of cognitive shortcoming: agents who *are* epistemically responsive to null events, but who fail to recognize the strategic implications of such responsiveness on the part of others, and who, like quintessential Watsonians, “overspeak.”

8 Appendix: Formal analysis

In this appendix, we consider the incentives perceived by Bayesian and by Watsonian agents in a variety of roles and settings. The scope of the analysis covers all of the situations faced by our experimental subjects. The solution concept that we employ is weak dominance.

Let $p(t)$ be the unconditional probability of type t . Member i 's prior beliefs about t_i are described by $p(t|a_i)$, which is derived from $p(t)$ and $\Pr(a_i|t) = \begin{cases} \frac{1}{2} & \text{if } a_i \in \{t^1, t^2\} \\ 0 & \text{otherwise} \end{cases}$ using Bayes Rule. Let M_i be the set of messages received by i during deliberation; then i 's posterior belief $p(t|a_i, M_i, \lambda)$ is also derived from $p(t|a_i)$ using Bayes Rule. Given that the mapping of type to ideal point is one-to-one and commonly known, in the interests of simplicity of reference, and with some abuse of notation, we use type and ideal point interchangeably (e.g., $p(\overline{AB})$ in place of $p((A, B))$).

Let $\mathcal{M}_i(a)$ be the set of all possible M_i given a . Note that M_i could contain two identical elements if i receives two identical messages. (The identity of the sender and the receiver is not observable.) We abuse set notation by writing, e.g., $M_i = \{B, B\}$ is i receives two “ B ” messages.

Let $v^*(p(\cdot))$ be i 's weakly dominant voting strategy. Given that $p(\overline{AB}|A) = 1$ and $p(\overline{CD}|B) = p(\overline{AB}|C) = 0$,

$$v_i^*(p(\cdot)) = \begin{cases} \overline{AB} & \text{if } p(\overline{AB}|a_i, M_i, \lambda) > p(\overline{BC}|a_i, M_i, \lambda) \\ \overline{BC} & \text{if } p(\overline{BC}|a_i, M_i, \lambda) > p(\overline{AB}|a_i, M_i, \lambda). \end{cases} \quad (1)$$

Let $\lambda_i^{BR} : \{0, 1\}^2 \rightarrow 2^{\{0,1\}}$ be i 's best response to λ_{-i} given v^* , i.e.,

$$\lambda_i^{BR} = \arg \max E[u_i(\lambda_i, \lambda_{-i}, v^*, a, p(t))].$$

8.1 Equilibrium Behavior of Bayesian Agents

Situation ABB , Vote \overline{AB} vs \overline{BC} . $\mathcal{M}_1((A, B, B)) = \{\emptyset, \{B\}, \{B, B\}\}$; $\mathcal{M}_2((A, B, B)) = \mathcal{M}_3((A, B, B)) = \{\emptyset, \{A\}, \{\text{foreign}\}, \{B\}, \{A, B\}, \{B, \text{foreign}\}\}$.

Suppose without loss of generality that $a_i = A$, $a_2 = a_3 = B$. From (1), if $p(\overline{BC}|B) > p(\overline{AB}|B)$, then $\forall j \in \{2, 3\}$, we have (1) $\forall M_j$ s.t. $A \notin M_j$, $v_j^*(p(\cdot|B, M_j, \lambda)) = \overline{BC}$; and (2) $\forall M_j \ni A$, $v_j^*(p(\cdot|B, M_j, \lambda)) = \overline{AB}$. If $p(\overline{BC}|A) < p(\overline{AB}|B)$, then $\forall j \in \{2, 3\}$, we have (1) $\forall M_j \ni \text{"foreign"}$, $v_j^*(p(\cdot|B, M_j, \lambda)) = \overline{BC}$; and (2) $\forall M_j$ s.t. "foreign" $\notin M_j$, $v_j^*(p(\cdot|B, M_j, \lambda)) = \overline{AB}$. Thus,

$$\lambda_1^{BR} = \begin{cases} \{0\} & \text{if } p(\overline{BC}) > p(\overline{AB}) \text{ and } \exists j \in \{2, 3\} \text{ s.t. } \lambda_j = 1 \\ \{1\} & \text{if } p(\overline{BC}) < p(\overline{AB}) \text{ and } \exists j \in \{2, 3\} \text{ s.t. } \lambda_j = 1 \\ \{0, 1\} & \text{if } \lambda_2 = \lambda_3 = 0 \end{cases}$$

$$\lambda_2^{BR} = \lambda_3^{BR} = \begin{cases} \{1\} & \text{if } \lambda_1 = 0 \\ \{0, 1\} & \text{if } \lambda_1 = 1, \end{cases}$$

and the unique weakly dominant strategies are

$$\lambda_1^* = \begin{cases} \{0\} & \text{if } p(\overline{BC}) > p(\overline{AB}) \\ \{1\} & \text{if } p(\overline{BC}) < p(\overline{AB}) \end{cases}$$

$$\lambda_2^* = \lambda_3^* = 1. \blacksquare$$

Situation ABC , Vote \overline{AB} vs \overline{BC} . $\mathcal{M}_1((A, B, C)) = \{\emptyset, \{B\}, \{\text{foreign}\}, \{B, \text{foreign}\}\}$; $\mathcal{M}_2((A, B, C)) = \{\emptyset, \{A\}, \{\text{foreign}\}, \{C\}, \{A, \text{foreign}\}, \{C, \text{foreign}\}\}$; $\mathcal{M}_3((A, B, C)) = \{\emptyset, \{B\}, \{\text{foreign}\}, \{B, \text{foreign}\}, \{\text{foreign}, \text{foreign}\}\}$.

Suppose without loss of generality that $a_i = A, a_2 = B, a_3 = C$. From (1), $v_1^*(p(\cdot)) = \overline{AB}$ and $v_3^*(p(\cdot)) = \overline{BC} \forall M, \forall \lambda$. Thus, $\lambda_2^* = 1$ is a unique weakly dominant strategy.

If $p(\overline{BC}|B) > p(\overline{AB}|B)$, then $v_2^*(p(\cdot|B, M_2, \lambda)) = \overline{AB}$ iff $M_2 \ni A$ or both $M_2 =$ “foreign” and $\lambda_1 = 1, \lambda_3 = 0$; else $v_2^*(p(\cdot|B, M_2, \lambda)) = \overline{BC}$. Thus $\lambda_3^* = 1$ is a unique weakly dominant strategy. $\lambda_1^{BR}(1, 1) = \{0\}$, thus $\lambda_1^* = 0$, given $\lambda_2^* = \lambda_3^* = 1$.

If $p(\overline{AB}|B) > p(\overline{BC}|B)$, then $v_2^*(p(\cdot|B, M_2, \lambda)) = \overline{BC}$ iff $M_2 \ni C$ or both $M_2 =$ “foreign” and $\lambda_1 = 0, \lambda_3 = 1$; else $v_2^*(p(\cdot|B, M_2, \lambda)) = \overline{AB}$. Thus $\lambda_1^* = 1$ is a unique weakly dominant strategy. $\lambda_3^{BR}(1, 1) = \{0\}$, thus $\lambda_3^* = 0$, given $\lambda_1^* = \lambda_2^* = 1$. ■

8.2 Equilibrium Behavior of Watsonian Agents

Note that the arguments identical to the ones below establish the identical optimal speaking/listening choices of the Unreflective Bayesian agents.

Situation ABB , Vote \overline{AB} vs \overline{BC} . Suppose without loss of generality that $a_i = A, a_2 = a_3 = B$. From (1), $v^*(p(\cdot|B, M_1, \lambda)) = \overline{AB} \forall M_1, \forall \lambda$. Because $v^*(p(\cdot|B, M_j, \lambda)) = v^*(p(\cdot|B, M_j \setminus B, \lambda)) \forall M_j, \forall \lambda$, $\lambda_2^* = \lambda_3^* = 1$ is the unique weakly dominant strategy.

If $p(\overline{BC}|A) < p(\overline{AB}|B)$, then $\forall j \in \{2, 3\}, v^*(p(\cdot|B, M_j, \lambda)) = \overline{AB} \forall M_j, \forall \lambda$. Thus $\lambda_1^{BR}(1, 1) = \{0, 1\}$, and $\lambda_1^* \in \{0, 1\}$.

If $p(\overline{BC}|A) > p(\overline{AB}|B)$, then $\forall j \in \{2, 3\}, v^*(p(\cdot|B, M_j, \lambda)) = \overline{AB}$ iff $M_j \ni A$; else $v^*(p(\cdot|B, M_j, \lambda)) = \overline{BC}$. Thus, $\lambda_1^* = 0$ is the unique weakly dominant strategy. ■

Situation ABC , Vote \overline{AB} vs \overline{BC} . Suppose without loss of generality that $a_i = A, a_2 = B, a_3 = C$. From (1), $v^*(p(\cdot|A, M_1, \lambda)) = \overline{AB} \forall M_1, \forall \lambda$ and $v^*(p(\cdot|C, M_3, \lambda)) = \overline{BC} \forall M_3, \forall \lambda$. Thus, $\lambda_2^* = 1$ is a unique weakly dominant strategy.

If $p(\overline{AB}|B) > p(\overline{BC}|B)$, then $v^*(p(\cdot|B, M_2, \lambda)) = \overline{BC}$ iff $M_2 \ni C$ else $v^*(p(\cdot|B, M_2, \lambda)) = \overline{AB}$. Thus $\lambda_3^* = 0$ is a unique weakly dominant strategy. $\lambda_1^{BR}(1, 0) = \{0, 1\}$, so $\lambda_1^* \in \{0, 1\}$.

If $p(\overline{BC}|B) > p(\overline{AB}|B)$, then $v^*(p(\cdot|B, M_2, \lambda)) = \overline{AB}$ iff $M_2 \ni A$ else $v^*(p(\cdot|B, M_2, \lambda)) = \overline{BC}$. Thus $\lambda_1^* = 0$ is a unique weakly dominant strategy. $\lambda_3^{BR}(0, 1) = \{0, 1\}$, thus $\lambda_3^* \in \{0, 1\}$. ■

References

- [1] Austen-Smith and Feddersen. 2005. "Deliberation, Preference Uncertainty, and Voting Rules." Northwestern University Mimeo.
- [2] Baron, David P. 2003. "Private Politics." *Journal of Economics & Management Strategy* 12 (1), 31-66.
- [3] Baron, Jonathan. 1994. *Thinking and Deciding*. Cambridge University Press.
- [4] Binmore, Ken. 1990. *Essays on the Foundations of Game Theory*. London: basil Blackwell.
- [5] Bohman, James and William Rehg, eds. 1997. *Deliberative Democracy: Essays on Reason and Politics*. MIT Press.
- [6] Calvert, Randall and James Johnson. 1998. "Rational Actors, Political Argument and Democratic Deliberation." University of Rochester Mimeo.
- [7] Cohen, Joshua. 1997. "Deliberation and Democratic Legitimacy." In J. Bohman and W. Rehg, eds., *Deliberative Democracy: Essays on Reason and Politics*. MIT Press, 67-92.
- [8] Crawford, Vincent and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50, 1431-51.
- [9] Doyle, Arthur Conan. 1976a. "Silver Blaze." In *The Complete Sherlock Holmes*. Doubleday, 330-49.
- [10] Doyle, Arthur Conan. 1976b. "The Sign of Four." In *The Complete Sherlock Holmes*. Doubleday, 88-160.
- [11] Dawes, Robyn M. 1998. "Behavioral Decision Making and Judgment." In D. T. Gilbert, et al. eds., *The Handbook of Social Psychology*, 4th ed. McGraw Hill, 497-548.
- [12] Elster, Jon. 1997. "The Market and the Forum: Three Varieties of Political Theory." In J. Bohman and W. Rehg, eds., *Deliberative Democracy*. Cambridge: MIT Press, 3-34.

- [13] Fearon, James. 1998. "Deliberation as Discussion." In Jon Elster, ed., *Deliberative Democracy*. Cambridge University Press.
- [14] Fischbacher, Urs. 1999. "z-Tree - Zurich Toolbox for Readymade Economic Experiments - Experimenter's Manual." Working Paper Nr. 21, Institute for Empirical Research in Economics, University of Zurich.
- [15] Fishkin, James, and Robert C. Luskin. 1996. "The Deliberative Poll: A Reply to our Critics," *Public Perspective* 7(1), 45-49.
- [16] Fishkin, James, Robert C. Luskin, and Roger Jowell. 2000. "Deliberative Polling and Public Consultation," *Parliamentary Affairs* 53, 657-666.
- [17] French, Damien, and Michael Laver. 2005. "Participation Bias and Framing Effects in Citizens' Juries." Paper, Annual Meeting of the American Political Science Association.
- [18] Gerardi, Dino and Leeat Yariv. 2002. "Putting Your Mouth Where Your Mouth Is: An Analysis of Collective Choice With Communication." Yale University Mimeo.
- [19] Glazer, Jacob and Ariel Rubinstein. 2005. "On the Pragmatics of Persuasion: a Game Theoretical Approach." Tel-Aviv University Mimeo.
- [20] Habermas, Jurgen. 1990. *Moral Consciousness and Communicative Action*. Cambridge: MIT Press.
- [21] Hafer, Catherine and Dimitri Landa. 2005. "Deliberation, Ideological Bias, and Group Choice." New York University Mimeo.
- [22] Hafer, Catherine and Dimitri Landa. 2006. "Deliberation as Self-Discovery and Institutions for Political Speech." *Journal of Theoretical Politics*, forthcoming.
- [23] Lipman, Bart and Duane J. Seppi. 1995. Robust Inference in Communication Games with Partial Proveability. *Journal of Economic Theory* 66, 370-405.

- [24] Lord, C. G., L. Ross, and M. R. Lepper. 1979. "Biased Assimilation and Attitude Polarization: The Effects of Prior Theories on Subsequently Considered Evidence." *Journal of Personality and Social Psychology* XXXVII, 2098-2109.
- [25] Lupia, Arthur. 2002. "Deliberation Disconnected: What It Takes to Improve Civic Competence." *Law and Contemporary Problems* 65 (3), 133-50.
- [26] Lupia, Arthur and Mathew D. McCubbins. 1998. *The Democratic Dilemma: Can Citizens Learn What They Need To Know?* Cambridge: Cambridge University Press.
- [27] Manin, Bernard. 1987. "On Legitimacy and Political Deliberation." *Political Theory* 15 (3), 338-68.
- [28] Meirowitz, Adam. 2005. "In Defense of Exclusionary Deliberation: Communication and Voting with Private Beliefs and Values." Princeton University Mimeo.
- [29] Patty, John. 2005. "Arguments-Based Collective Choice." Harvard University Mimeo.
- [30] Rabin, Matthew. 1998. "Psychology and Economics." *Journal of Economic Literature* XXXVI (March), 11-46.
- [31] Rabin, Matthew and Joel L. Schrag. 1999. "First Impressions Matter: A Model of Confirmatory Bias." *The Quarterly Journal of Economics* (February), 37-82.
- [32] Schotter, Andrew. 2005. "Strong and Wrong: On The Use of Rational Choice Theory in Experimental Economics." NYU Department of Economics Mimeo.
- [33] Stasavage, David. 2005. "Polarization and Public Deliberation." LSE Working Paper.
- [34] Wason, P. C. 1977. "Self-Contradictions." In P. N. Johnson-Laird and P. C. Wason, eds., *Thinking: Readings in Cognitive Science*. Cambridge University Press, 114-28.
- [35] Wason, P. C. 1968. "Reasoning About a Rule." *Quarterly Journal of Experimental Psychology* 20, 273-81.
- [36] Zaller, John. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Table 1a. Weakly Dominant Deliberative Strategies for Bayesian Agents

Deliberative Setting	A active (Extreme)	B active (Moderate)	C active (Extreme)
$ABB, Pr(\overline{BC} B) > Pr(\overline{AB} B)$	Speak	Listen	-
$ABB, Pr(\overline{BC} B) < Pr(\overline{AB} B)$	Listen	Listen	-
$ABC, Pr(\overline{BC} B) > Pr(\overline{AB} B)$	Speak	Listen	Listen
$ABC, Pr(\overline{BC} B) < Pr(\overline{AB} B)$	Listen	Listen	Speak

Note: The weakly dominant deliberative strategies for agents with an ‘A’ or ‘C’ active fragment are indicated in plain text for Case I situations, and in boldface for Case II situations.

Table 1b. Dominant Voting Strategies for Bayesian Agents with a ‘B’ (Moderate) Active Fragment, by Communications Received

Deliberative Setting	nothing	foreign fragment only	latent fragment
$ABB, Pr(\overline{BC} B) > Pr(\overline{AB} B)$	\overline{BC}	\overline{BC}	actual true number
$ABB, Pr(\overline{BC} B) < Pr(\overline{AB} B)$	\overline{AB}	\overline{BC}	actual true number
$ABC, Pr(\overline{BC} B) > Pr(\overline{AB} B)$	\overline{BC}	\overline{BC}	actual true number
$ABC, Pr(\overline{BC} B) < Pr(\overline{AB} B)$	\overline{AB}	\overline{AB}	actual true number

Note: subjects with an ‘A’ (‘C’) active fragment always have a dominant strategy to vote for \overline{AB} over \overline{BC} (\overline{BC} over \overline{AB}).

Table 2a. Aggregate Communications Behavior (Periods 13-30)

Deliberative Setting	A active (Extreme)	B active (Moderate)	C active (Extreme)
$ABB, Pr(\overline{BC} B) > Pr(\overline{AB} B)$	79.2% Speak (38/48)	99.0% Listen (95/96)	-
$ABB, Pr(\overline{BC} B) < Pr(\overline{AB} B)$	54.2% Speak (26/48)	97.9% Listen (94/96)	-
$ABC, Pr(\overline{BC} B) > Pr(\overline{AB} B)$	78.3% Speak (47/60)	96.7% Listen (58/60)	58.3% Speak (35/60)
$ABC, Pr(\overline{BC} B) < Pr(\overline{AB} B)$	66.7% Speak (40/60)	98.3% Listen (59/60)	73.3% Speak (44/60)

**Table 2b. Aggregate Communications Behavior When Subjects Have ‘A’ or ‘C’
(Extreme) Active Fragments**

.	Speak (Case I)	Listen (Case I)	Speak (Case II)	Listen (Case II)
Experimental Data				
All Periods	196 (71.0%)	80 (29.0%)	169 (61.2%)	107 (38.8%)
Periods 1-12	67 (62.0%)	41 (38.0%)	68 (63.0%)	40 (37.0%)
Periods 13-30	129 (76.8%)	39 (23.2%)	101 (60.1%)	67 (39.9%)
Theoretically Expected Behavior				
Bayesian	100%	0%	0%	100%

**Table 3. Factors Affecting Deliberative Choice in Subjects with ‘A’ or ‘C’
(Extreme) Active Fragments (Periods 13-30)**

Probit Regression with Robust Standard Errors

Dependent Variable: listen = 1 if subject listens, = 0 if subject speaks

$N = 336$; $PseudoR^2 = 0.0321$

listen	coefficient	robust SE	z	$P > z $
lnperiod	-.1352728	.3447588	-0.39	0.695
caselisten	.4685002**	.1472395	3.18	0.001
<i>ABB</i> (not <i>ABC</i>)	.0563819	.1722589	0.33	0.743
right-handed	.2296582	.1752874	1.31	0.190
xdist	.0100337	.0216525	0.46	0.643
xfrac	-.4254516	1.569162	-0.27	0.786
pdist	-1.562123	2.172081	-0.72	0.472
pfrac	.7623341	1.506065	0.51	0.613
constant	-.4361045	1.111846	-0.39	0.695

lnperiod is the natural log of the period number. **caselisten** is 1 for Case II, 0 for Case I. **ABB (not ABC)** is 1 for ABB or CCD, 0 for ABC or BCD. **right-handed** is 1 for BCD or CCD, 0 for ABB or ABC. **xdist** is $|y_1 - y_2|$, the distance between the alternatives to be voted on. **xfrac** is $\frac{|y_1 - y_2|}{CD - AB}$, the distance between the alternatives to be voted on relative to the distance between the most extreme true numbers. **pdist** is the absolute value of the difference of the unconditional probabilities associated with the alternatives to be voted on. **pfrac** is **pdist** divided by the sum of the unconditional probabilities associated with the alternatives to be voted on.

Table 4a. Aggregate Voting Behavior of Subjects with a ‘B’ (Moderate) Active Fragment (All Periods)

Deliberative Setting	nothing	foreign fragment only	latent fragment
$ABB, Pr(\overline{BC} B) > Pr(\overline{AB} B)$	91.7% \overline{BC} (44/48)	91.8% \overline{BC} (67/73)	92.7% actual true number (38/41)
$ABB, Pr(\overline{BC} B) < Pr(\overline{AB} B)$	94.6% \overline{AB} (70/74)	56.3% \overline{BC} (18/32)	100% actual true number (59/59)
$ABC, Pr(\overline{BC} B) > Pr(\overline{AB} B)$	100% \overline{BC} (18/18)	95.5% \overline{BC} (21/22)	96.4% actual true number (54/56)
$ABC, Pr(\overline{BC} B) < Pr(\overline{AB} B)$	100% \overline{AB} (18/18)	85.7% \overline{AB} (12/14)	90.6% actual true number (58/64)

Note: subjects with an ‘A’ active fragment voted for \overline{AB} , their dominant strategy, 96.7% of the time (348/360). Subjects with a ‘C’ active fragment voted for \overline{BC} , their dominant strategy, 96.4% of the time (185/192).

Table 4b. Voting Behavior of Subjects with a ‘B’ (Moderate) Active Fragment Who Receive Only a Foreign Fragment

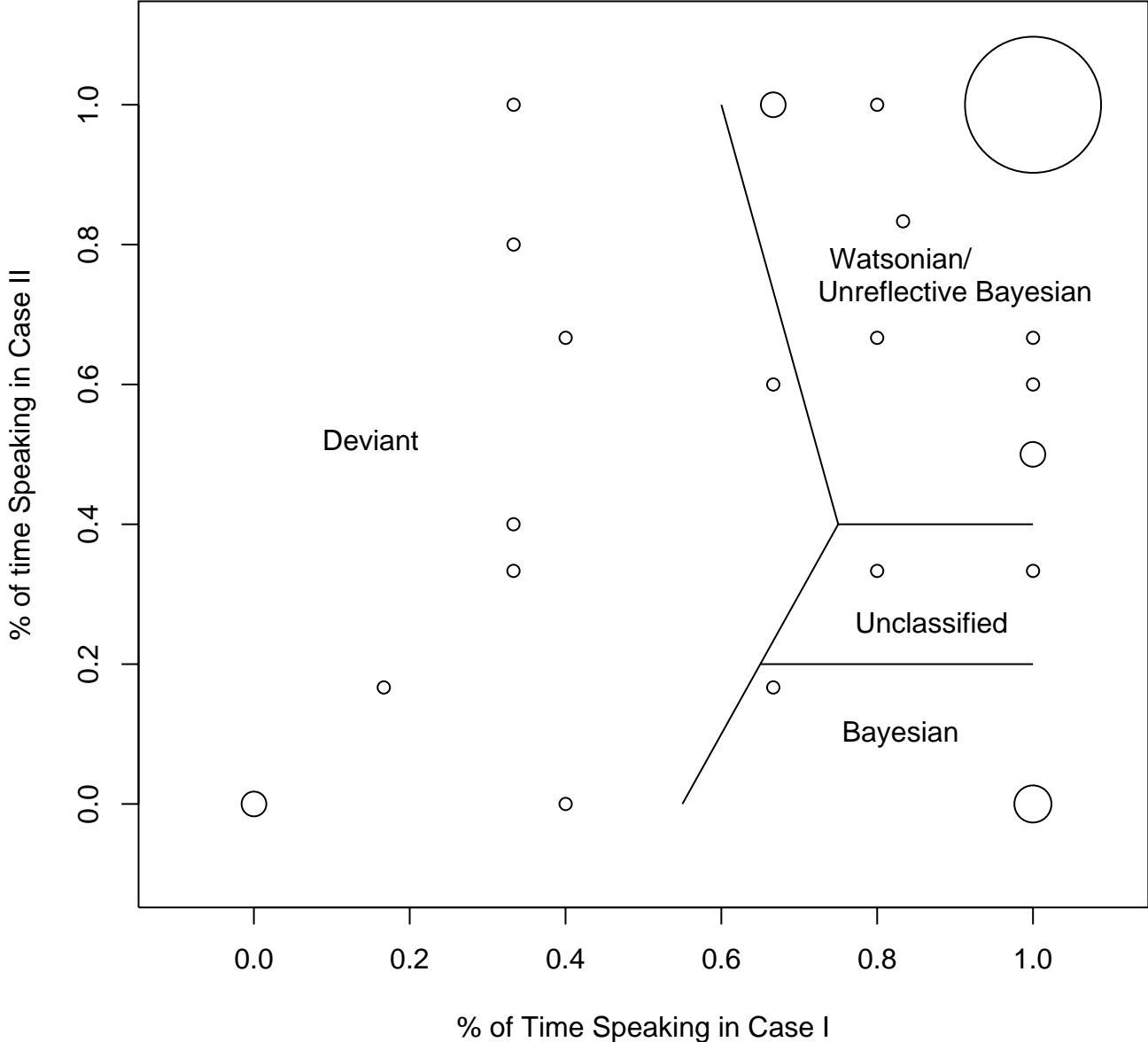
Case	Voted Correctly in Case I or Case II (ABC)	Voted Correctly in Case II (ABB)
All periods	91.7% (100/109) (with prior)	56.3% (18/32) (against prior)
Periods 13-30	93.4% (71/76) (with prior)	62.5% (10/16) (against prior)

**Table 5. Individual Communications Behavior When Subjects Have ‘A’ or ‘C’
(Extreme) Active Fragments (Periods 13-30)**

Subject No.	Speak (Case I)	Listen (Case I)	Speak (Case II)	Listen (Case II)	Classification
1	0	3	0	5	Dev.
2	4	1	1	2	Unc.#
3	6	0	3	3	W/UB#
4	4	1	2	1	W/UB
5	5	0	3	0	W/UB*
6	3	0	0	5	B*
7	1	2	2	3	Dev.*
8	6	0	0	6	B
9	2	4	2	4	Dev.
10	1	5	1	5	Dev.#
11	4	2	1	5	Lean. B**
12	5	0	3	0	W/UB#
13	2	4	6	0	Dev.
14	3	0	5	0	W/UB#
15	3	0	3	2	W/UB
16	5	0	3	0	W/UB*
17	2	1	3	2	Dev.
18	5	0	3	0	W/UB
19	3	0	5	0	W/UB*
20	5	0	3	0	W/UB*
21	5	1	5	1	W/UB*
22	2	3	0	3	Dev.
23	5	0	3	0	W/UB###
24	3	0	5	0	W/UB**
25	3	0	5	0	W/UB*
26	6	0	3	3	W/UB*
27	6	0	0	6	B*
28	6	0	6	0	W/UB#
29	6	0	4	2	W/UB*
30	5	0	1	2	Unc.*
31	4	2	6	0	Lean. W/UB
32	1	2	4	1	Dev.*
33	0	3	0	5	Dev.###
34	2	3	2	1	Dev.##
35	2	1	5	0	W/UB
36	4	1	3	0	W/UB#*
TOTALS	129 (76.8%)	39 (23.2%)	101 (60.1%)	67 (39.9%)	.

Dev. - Deviant; *Unc.* - Unclassified; *W* - Watsonian; *UB*- Unreflective Bayesian; *B* - Bayesian; *Lean. B* - Leaning Bayesian; *Lean. W* - Leaning Watsonian. The votes of those subjects who in any period of the experiment received only a foreign fragment while possessing active fragment *B* in Case II (*ABB*) are indicated next to their classification type. Each occurrence of (#) denotes one *incorrect* vote *for* the prior while each occurrence of (*) denotes one *correct* vote *against* the prior in such a situation.

Figure 1. Experimental Data on Deliberative Behavioral Types.



SUPPLEMENTAL APPENDIX FOR REVIEWERS: INSTRUCTIONS TO SUBJECTS

Instructions

Introduction

This is an experiment on decision-making. In the following experiment you will make a series of choices. At the end of the experiment, you will be paid depending on the specific choices that you made during the experiment and the specific choices made by other people. If you follow the instructions and make appropriate decisions, you may make an appreciable amount of money.

The experiment will consist of 30 different rounds. At the beginning of each round of the experiment, you will be randomly assigned into a group of three people (that is, you will be randomly grouped with two other people in the lab). As such, the composition of groups will be different in each round of the experiment. All of your interactions with others will take place anonymously through a computer terminal, so that your personal identity will never be revealed to others and you will not know who is in your group.

In each round, each member of your group will have what we call a “true number,” and will be given partial information about it. Each member of your group will, then, have an opportunity to communicate with the other members. At the end of each round, each group member will be asked to cast a vote in an election between two different two-digit numbers. Ultimately, you will make the most money if the number that wins the majority of votes is as close as possible to your “true number.”

The structure of the experiment is now described in greater detail.

(1) Initial Information

At the beginning of each round of the experiment, you will receive four kinds of information.

First, you will be given the set of possible true numbers. Such a set might look as follows: {24, 47, 79}. One of these numbers will be your own true number, although you will not be told which of these numbers it is. Similarly, each of the other members of your group will also have a true number that comes from this set, and will be shown the set of numbers but will not be told what his or her particular true number is. It may be the case that every member of your group has a different true number, but it may also be that some of the members of your group have the same true number. Thus, the true number of any other member of your group may be the same as your true number, or it may be different.

A second piece of information that is given to everyone in your group will further help you figure out your true number. For each of the possible true numbers, you will be told how likely it would be, *if you didn't know anything else about what your true number was*, that a given number in the set of true numbers would, in fact, turn out to be *your* true number. For example, you may be told that that likelihood is 20% for number 24, 50 % for number 47, and 30% for number 79.

But, in fact, we will give you an additional hint. To help you further improve your guesses of your own true numbers, each of you will receive a third piece of information – a one-digit fragment of your own true number. For example, if your true number were 47, you might be given the fragment “4” or the fragment “7.” At the same time that you receive this fragment of your own true number, you will also be told the set of one-digit fragments received by all the other members of your group.

Finally, you will be told which alternatives will be voted on in the election held at the end of the round. For example, you may be told that the election will be between the number 24 and the number 47.

(2) Communication

After you have received the initial information, all of you will have the opportunity to communicate with other members of your group.

Communication in your group will take a particular form. All members of your group will, simultaneously, be asked to choose either to (a) send their respective fragments to all the other group members at once; *or* to (b) receive the fragments that are being sent by all those members of your group who choose to send their fragments (that is, by all the members of your group who choose the option (a)). Each member of the group must choose either to send or to receive; it is not possible to do both.

If you choose to *receive* fragments, you will receive fragments from those and only those members of your group who choose to send their fragments. Regardless of how many fragments you may receive, you will only be able to see a fragment sent to you if it is part of your true number – i.e., if it is a fragment that completes your true number or a fragment you already have. If you receive other fragments – that is, fragments that are not part of your true number – you will only be told that you have received “foreign” fragments, and how many of them. For example, suppose that your fragment is 4 and your true number is 47. If you choose to receive fragments, and a member of your group sends the fragment 7, you will be told that you have received the fragment 7. Similarly, if you choose to receive fragments, and a member of your group sends the fragment 4, you will be told that you have received the fragment 4. However, if you choose to receive fragments, and a member of your group sends the fragment 9, you will simply be told that you have received a “foreign” fragment, because 9 is not a fragment of 47. Likewise, every other member of your group who chooses to receive fragments will see only those fragments that are a part of *his or her own* true number, and will be told how many “foreign” fragments they have received in the event that any such fragments are received.

If you choose to *send* your fragment, it will be received by those and only those members of your group who choose to receive fragments, but you yourself will receive no fragments of any kind. Similarly, if other members of your group choose to send their respective fragments, these fragments will be received by those and only those members who choose to receive, but the sending members themselves will receive no fragments of any kind.

Communication may then have two effects: First, you might learn something about your own true number that could help you determine how you will vote. And second, you might cause other members of your group to change their guesses about their own true numbers and so

affect *their* votes. Always remember that you will make more money the closer the number that wins the majority of votes is to your true number.

(3) Election and Payoffs

Once the communication phase is complete, all members of your group will be asked to vote in the election that was specified earlier – in the example above, this was 24 vs. 47. The number that receives the majority of votes wins. Because there are three members in each group (including you), there cannot be a tie. The winning number from the election will be used to determine the payoffs for the round.

Your payoffs for each round will be calculated as follows. If the winning number matches your true number exactly, you will receive 80 cents. If the winning number does not match your true number exactly, you will receive 80 cents less 1 cent for each “unit of distance” between the winning number and your true number. For example, if your true number is 62, but the winning number is 12, there are 50 “units of distance” between your true number and the winning number, and you would receive $(80 - 50)$ cents, that is, 30 cents.

The experiment will consist of 30 rounds like the one just described. In each of these rounds, you will be assigned a new true number; receive initial information about the new set of possible true numbers; be told what numbers are to be voted upon; be given a fragment of your true number; have an opportunity to attempt communication with other members of your group; and, along with your fellow group members, cast a vote. Your total payoff for the experiment is the sum of your payoffs from each of these rounds, plus the show-up fee. Remember, your payoff in each round is higher, the closer the winning number is to your true number.