Theoretical Economics 1 (2006), 395–410 1555-7561/20060395

A study in the pragmatics of persuasion: a game theoretical approach

JACOB GLAZER
Faculty of Management, Tel Aviv University and Department of Economics, Boston University

ARIEL RUBINSTEIN
School of Economics, Tel Aviv University and Department of Economics, New York University

A speaker wishes to persuade a listener to take a certain action. The conditions under which the request is justified, from the listener's point of view, depend on the state of the world, which is known only to the speaker. Each state is characterized by a set of statements from which the speaker chooses. A persuasion rule specifies which statements the listener finds persuasive. We study persuasion rules that maximize the probability that the listener accepts the request if and only if it is justified, given that the speaker maximizes the probability that his request is accepted. We prove that there always exists a persuasion rule involving no randomization and that all optimal persuasion rules are ex-post optimal. We relate our analysis to the field of pragmatics.

KEYWORDS. Persuasion, mechanism design, hard evidence, pragmatics.

JEL CLASSIFICATION. C61, D82, D83.

1. INTRODUCTION

A persuasion situation involves an agent (*the speaker*) who attempts to persuade another agent (*the listener*) to take a certain action. Whether or not the listener should accept the speaker's suggestion depends on information possessed by the speaker. In such a situation, the speaker often presents hard evidence to support his position, but is restricted as to how many pieces of evidence he can present. This restriction may be due either to time constraints or to limitations on the listener's capability to process information. Our purpose in this paper is to shed light on the rules that determine which of the facts, presented by the speaker, the listener will find persuasive.

The topic of this paper is related to a field in linguistics called pragmatics, which explores the rules that determine how people interpret an utterance, made in the course of a conversation, beyond its literal content (see Grice 1989). Grice suggested that the leading principle in the interpretation of utterances is what he termed the "cooperative principle", according to which the interpretation of utterances in a regular conversation

Jacob Glazer: glazer@post.tau.ac.il
Ariel Rubinstein: rariel@post.tau.ac.il
We are grateful to an editor of this journal and anonymous referees for their thoughtful comments.

can be made on the assumption that the speaker and the listener have common interests. However, the cooperative principle does not appear to be relevant in a persuasion situation in which the agents may have conflicting interests.

The following example clarifies the distinction between the pragmatics of conversation and the pragmatics of persuasion: You are discussing the chances of each of two candidates in an upcoming election. The electorate consists of ten voters. Assume that the other person has access to the views of these ten voters. Imagine that he has just informed you that a, d, and g support candidate A. If it is a friendly conversation, then you are most likely to think that he has selected three people who represent the views of the majority of the voters. Thus, you are likely to be persuaded that A is likely to win the election. If, on the other hand, independently of the truth, the other person is trying to persuade you that A will win, you will find this very same statement to be a weak argument since you will suspect that he has intentionally selected three supporters of A.

What governs the pragmatic rules of persuasion? We propose an approach analogous to Grice's cooperative principle in which the pragmatic rules of persuasion are determined by a fictitious designer before the discourse begins. These rules govern the speaker's choice of facts to present in the knowledge that the listener will interpret his statements according to these rules. The rules are structured by the designer to maximize the probability that the listener will make the "right" decision (from his point of view and given the "true" situation) on the basis of the information provided to him by a self-interested speaker and subject to constraints on the amount of information that can be submitted to him by the speaker.

We conduct our investigation within the narrow boundaries of a particular model in which several assumptions admittedly play a critical role. Our analysis is faithful to economic tradition rather than to the methodology of Pragmatics. Nevertheless, we believe that the study conducted here demonstrates the potential of research to find a uniform principle that guides individuals in interpreting statements in persuasion situations.

This paper belongs to a research program in which we apply a game theoretical approach to issues in Pragmatics. In Glazer and Rubinstein (2001) we study an example of a debate situation involving two parties each of whom tries to persuade a third party to accept his position. Even closer to this paper is Glazer and Rubinstein (2004), which analyzes a persuasion situation in which after the speaker makes his case the listener can obtain partial information about the state of the world. After specifying our current model, we will compare it to the one in Glazer and Rubinstein (2004).

2. The model

A speaker wishes to persuade a listener to take a certain action. The listener can either accept or reject the speaker's suggestion (there is no partial acceptance). Whether or not the listener should be persuaded depends on the *state*, which is an element in a set X. A set $A \subset X$ consists of all the states in which the listener would wish to be persuaded (i.e. to accept the speaker's suggestion) if he knew the state, and the set $R = X \setminus A$ consists of all the states in which the listener would wish to reject the speaker's request. The listener's initial beliefs about the state are given by a probability measure p over X. Denote by p_x the probability of state x.

We assume that for every state x, there is a set of statements $\sigma(x)$ that the speaker can make. Let $S = \cup_{x\in X}\sigma(x)$. The meaning of "making statement s" is to present proof that the event $\sigma^{-1}(s) = \{x \mid s \in \sigma(x)\}$ has occurred.

In state x the speaker can make one and only one of the statements in $\sigma(x)$. Thus, for example, if the speaker can choose between remaining silent, making the statement α, making the statement β, or making both statements, the set $\sigma(x)$ consists of the four elements *silence*, α, β, and $\alpha \wedge \beta$.

To summarize, we model a *persuasion problem* as a four-tuple $\langle X, A, p, \sigma\rangle$. We say that the persuasion problem is finite if X is finite. We refer to the pair $\langle X, \sigma\rangle$ as a *signal structure*.

COMMENT. We say that a signal structure $\langle Y, e\rangle$ is *vectoric* if Y is a product set, i.e. $Y = \times_{k\in K} Y_k$ for some set K and some sets Y_k, $k \in K$, and the speaker in state x can make a statement concerning the value of one of the components of x, that is, $e(x) = \{(k, v) \mid k \in K \text{ and } v = x_k\}$.

One might think that we could make do by analyzing only vectoric signal structures. To see that this is not the case, let $\langle X, \sigma\rangle$ be a signal structure. Let $\langle Y, e\rangle$ be the vectoric signal structure with $Y = \{0, 1\}^S$. Every state $x \in X$ can be represented by the vector $\varphi(x) \in Y$, which indicates the statements available at x, that is, $\varphi(x)(s) = 1$ if $s \in \sigma(x)$ and 0 otherwise. However, the two structures are not equivalent. First, we allow for the possibility that two states have the same set of feasible statements. Second, and more importantly, in the corresponding vectoric structure the speaker in any state is able to show the value of the component that corresponds to any statement s. In other words, he is always able to prove whether s is available or not. In contrast, in our framework the fact that the speaker can make the statement s does not necessarily mean that he can make a statement that proves that s is not available.

We have in mind a situation in which the speaker makes a statement and the listener must then either take the action a, thus accepting the speaker's position, or the action r, thus rejecting it. A persuasion rule determines how the listener responds to each of the speaker's possible statements. We define a *persuasion rule* f as a function $f : S \to [0, 1]$. The function f specifies the speaker's beliefs about how the listener will interpret each of his possible statements. The meaning of $f(s) = q$ is that following a statement s, with probability q the listener is "persuaded" and chooses a, the speaker's favored action. We call a persuasion rule f *deterministic* if $f(s) \in \{0, 1\}$ for all $s \in S$.

We assume that the speaker wishes to maximize the probability that the listener is persuaded. Thus, given a state x, the speaker solves the problem $\max_{s\in\sigma(x)} f(s)$. The value of the solution, denoted by $\alpha(f, x)$, is the maximal probability of acceptance that the speaker can induce in state x. For the case in which $\sigma(x)$ is infinite, the solution can be approached but is not attainable and therefore we define $\alpha(f, x) = \sup_{s\in\sigma(x)} f(s)$.

Given the assumption that the speaker maximizes the probability of acceptance, we define the (listener's) error probability $\mu_x(f)$ in state x as follows: If $x \in A$, then $\mu_x(f) = 1 - \alpha(f, x)$, and if $x \in R$, then $\mu_x(f) = \alpha(f, x)$. The *error probability* induced by the persuasion rule f is $m(f) = \sum_{x\in X} p_x\mu_x(f)$. Given a problem $\langle X, A, p, \sigma\rangle$, an *optimal* persuasion rule is one that minimizes $m(f)$.

Note that persuasion rules are evaluated according to the listener's interests while those of the speaker are ignored. In addition, we assume that all errors are treated symmetrically. Our analysis remains the same if we add a variable c_x for the (listener's) "costs" of an error in state x and define the objective function to minimize $\sum_{x\in X} p_x c_x \mu_x(f)$.

EXAMPLE 1 ("The majority of the facts supports my position"). There are five independent random variables, each of which takes the values 1 and 0 each with probability 0.5. A realization of 1 means that the random variable supports the speaker's position. The listener would like to accept the speaker's position if and only if at least three random variables take the value 1. In the process of persuasion, the speaker can present the realization of at most m random variables that support his position.

Formally, $X = \{(x_1,\ldots,x_5) \mid x_k \in \{0,1\} \text{ for all } k\}$, $A = \{x \mid n(x) \geq 3\}$ where $n(x) = \sum_k x_k$, $p_x = \frac{1}{32}$ for all $x \in X$, and $\sigma(x) = \{\kappa \mid \kappa \subseteq \{k \mid x_k = 1\} \text{ and } |\kappa| \leq m\}$.

If $m = 3$, the optimal persuasion rule states that the listener is persuaded if the speaker presents any three random variables that take the value 1. The more interesting case is $m = 2$. If the listener is persuaded by the presentation of any two random variables that support the speaker's position, then the error probability is $\frac{10}{32}$. The persuasion rule according to which the listener is persuaded only by the speaker presenting a set of two "neighboring" random variables ($\{1,2\}$, $\{2,3\}$, $\{3,4\}$, or $\{4,5\}$) with the value 1 reduces the error probability to $\frac{5}{32}$ (an error in favor of the speaker occurs in the four states in which exactly two neighboring random variables support the speaker's position and in the state $(1,0,1,0,1)$ in which the speaker is not able to persuade the listener to support him even though he should).

The two mechanisms above do not use lotteries. Can the listener do better by applying a random mechanism? What is the optimal mechanism in that case? We return to this example after presenting some additional results. ◊

COMMENT. At this point, we wish to compare the current model with the one studied in Glazer and Rubinstein (2004). Both models deal with a persuasion situation in which (a) the speaker attempts to persuade the listener to take a particular action and (b) only the speaker knows the state of the world and therefore whether or not the listener should accept the speaker's request.

Unlike the current model, the speaker in the previous model could first send an arbitrary message (cheap talk) to the listener. After receiving the message, the listener could ask the speaker to present some hard evidence to support his request. The state of the world in that model is a realization of two random variables and the listener is able to ask the speaker to reveal at most one of them. Thus, unlike the current model, in which the speaker simply decides which hard evidence to present, in the previous model the speaker has to "follow the listener's instructions" and the listener can apply a random device to determine which hard evidence he asks the speaker to present. That randomization was shown to often be a critical element in the listener's optimal persuasion rule (a point further discussed below). On the other hand, in the previous model we do not allow randomization during the stage in which the listener finally decides whether or

not to accept the speaker's request, which we do allow in the current model. Allowing for such randomization in the previous model, however, is not beneficial to the listener, as we show to be the case in the current paper as well.

The randomization in the previous paper is employed during the stage in which the listener has to decide which hard evidence to request from the speaker. Note that if in that model we restrict attention to deterministic persuasion rules, then it is a special case of the current model. Eliminating randomization on the part of the listener in order to verify the information presented by the speaker, allows us to think about the persuasion situation in the previous model as one in which the speaker chooses which hard evidence to present rather than one in which the listener chooses which hard evidence to request.

Randomization plays such an important role in the previous model because it is, in fact, employed as a verification device. Without randomization, there is no value to the speaker's message since he could be lying. The listener uses randomization to induce the speaker to transfer more information than the information that is eventually verified.

Although the current model draws some inspiration from the previous one, the two papers relate to different persuasion situations and the results of the current paper cannot be derived from those of the previous one.

3. Two lemmas

We now present two lemmas that are useful in deriving an optimal persuasion rule.

3.1 *A finite number of persuasive statements is sufficient*

Our first observation is rather technical though simple. We show that if the set of states X is finite then even if the set of statements S is infinite there is an optimal persuasion rule in which at most $|X|$ statements are persuasive with positive probability.

Lemma 1. *Let $\langle X, A, p, \sigma\rangle$ be a finite persuasion problem.*

(i) An optimal persuasion rule exists.

(ii) There is an optimal persuasion rule in which $\{s \mid f(s) > 0\}$ does not contain more than $|X|$ elements.

Proof. Consider a partition of S such that s and s' are in the same cell of the partition if $\sigma^{-1}(s) = \sigma^{-1}(s')$. This partition is finite. Let T be a set of statements consisting of one statement from each cell of the partition. We now show that for every persuasion rule f, there is a persuasion rule g that takes a positive value only on T, such that $\alpha(g,x) = \alpha(f,x)$ for all x and thus $m(g) = m(f)$.

For every $s \in T$ let S_s be the cell in the partition of S that contains s. Define $g(s) = \sup_{s' \in S_s} f(s')$. For every $s \notin T$ define $g(s) = 0$.

For every state x,

$$\alpha(g,x)=\max_{s\in T\cap\sigma(x)} g(s)=\max_{s\in T\cap\sigma(x)}\sup_{s'\in S_s} f(s')=\sup_{s'\in\sigma(x)} f(s')=\alpha(f,x).$$

Thus, we can confine ourselves to persuasion rules that take the value 0 for any statement besides those in the finite set T. Any such persuasion rule is characterized by a vector in the compact set $[0,1]^T$. The error probability is a continuous function on this space and thus there is an optimal persuasion rule f^* with $f^*(s)=0$ for all $s\notin T$.

For every $x\in S$ let $s(x)\in\sigma(x)$ be a solution of $\max_{s\in\sigma(x)} f^*(s)$. Let g^* be a persuasion rule such that

$$g^*(s)=\begin{cases} f^*(s) & \text{if } s=s(x) \text{ for some } x\\ 0 & \text{otherwise.}\end{cases}$$

The persuasion rule g^* is optimal as well since $\alpha(g^*,x)=\alpha(f^*,x)$ for all x and thus $m(g^*)=m(f^*)$. Thus, we can confine ourselves to persuasion rules for which the number of statements that persuade the listener with positive probability is no larger than the size of the state space. □

3.2 *The "L-principle"*

The following result is based on an idea discussed in Glazer and Rubinstein (2004).

Let $\langle X,A,p,\sigma\rangle$ be a persuasion problem such that for all $x\in X$, $\sigma(x)$ is finite. We say that a pair (x,T), where $x\in A$ and $T\subseteq R$, is an L if for any $s\in\sigma(x)$ there is $t\in T$ such that $s\in\sigma(t)$. That is, an L consists of an element x in A and a set T of elements in R such that every statement that can be made by x can also be made by some member of T. An L, (x,T) is minimal if there is no $T'\subset T$ such that (x,T') is an L.

LEMMA 2 (The L-Principle). *Let (x,T) be an L in the persuasion problem $\langle X,A,p,\sigma\rangle$ and let f be a persuasion rule. Then $\sum_{t\in\{x\}\cup T}\mu_t(f)\geq 1$.*

PROOF. Recall that $\mu_x(f)=1-\alpha(f,x)$ and for every $t\in T$, $\mu_t(f)=\alpha(f,t)$. Therefore,

$$\sum_{t\in\{x\}\cup T}\mu_t(f)\geq\mu_x(f)+\max_{t\in T}\mu_t(f)\geq\mu_x(f)+\max_{s\in\sigma(x)} f(s)=\mu_x(f)+\alpha(f,x)=1.$$ □

The following example demonstrates how the L-principle can be used to verify that a certain persuasion rule is optimal. For any persuasion problem, the L-principle provides a lower bound on the probability of error that can be induced by a persuasion rule. Thus, if a particular persuasion rule induces a probability of error equal to a lower bound derived from the L-principle, then one can conclude that this persuasion rule is optimal.

EXAMPLE 2 ("I have outperformed the population average"). Consider a situation in which a speaker wishes to persuade a listener that his average performance in two previous tasks was above the average performance of the population. Denote by x_1 the proportion of the population that performed worse than the speaker in the first task

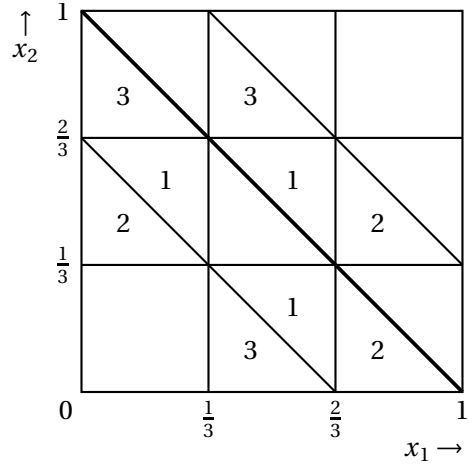

FIGURE 1. An optimal persuasion rule for Example 2.

and by x_2 the proportion of the population that performed worse than the speaker in the second task. The speaker wishes to persuade the listener that $x_1 + x_2 \geq 1$. The speaker knows his relative performance in the two tasks (that is, he knows x_1 and x_2) but can present details of his performance in only one of the tasks. We assume that the speaker's performances in the two tasks are uncorrelated. Formally, the signal structure is vectoric with $X = [0,1] \times [0,1]$; the probability measure p is uniform on X; and $A = \{(x_1, x_2) \mid x_1 + x_2 \geq 1\}$.

Note that if a statement is interpreted by the listener based only on its content, i.e. by stating that his performance was above $\frac{1}{2}$ in one of the tasks, the speaker persuades the listener and the probability of error is $\frac{1}{4}$.

The following argument (borrowed from Glazer and Rubinstein 2004) shows that there exists an optimal persuasion rule according to which the listener is persuaded by the speaker if and only if the speaker can show that his performance in one of the two tasks was above $\frac{2}{3}$. Furthermore, the minimal probability of error is $\frac{1}{6}$.

A minimal L in this case is any pair $(x, \{y, z\})$ where $x \in A$, y, $z \in R$, $x_1 = y_1$, and $x_2 = z_2$.

The set $T_1 = \{(x_1, x_2) \in A \mid x_1 \leq \frac{2}{3} \text{ and } x_2 \leq \frac{2}{3}\}$ is one of the three triangles denoted in Figure 1 by the number 1. Any three points $x = (x_1, x_2) \in T_1$, $y = (x_1 - \frac{1}{3}, x_2) \in R$ and $z = (x_1, x_2 - \frac{1}{3}) \in R$ establish an L. By the L-principle, for any persuasion rule f we have $\mu_x(f) + \mu_y(f) + \mu_z(f) \geq 1$. The collection of all these L's is a set of disjoint sets whose union is the three triangles denoted in the figure by the number 1. Therefore, the integral of $\mu_x(f)$ over these three triangles must be at least the size of T_1, namely $\frac{1}{18}$. Similar considerations regarding the three triangles denoted by the number 2 and the three triangles denoted by the number 3 imply that the minimal error probability is at least $\frac{1}{6}$. This error probability is attained by the persuasion rule according to which the listener is persuaded if and only if the speaker shows that either x_1 or x_2 take a value of at least $\frac{2}{3}$. ◊

4. Randomization is not needed

The next question to be addressed is whether randomization has any role in the design of the optimal persuasion rule. In other words, can the listener ever do better by making the speaker uncertain about the consequences of his statement? Glazer and Rubinstein (2004) show that in persuasion situations in which the listener can acquire partial information about the state of the world, uncertainty regarding what information he will acquire can be a useful device to the listener. However, as stated in Proposition 1 below, uncertainty is not useful to the listener in the present context.

Proposition 1. *(i) For every finite persuasion problem $\langle X, A, p, \sigma\rangle$, there exists an optimal persuasion rule f that is deterministic.*

(ii) For every persuasion problem $\langle X, A, p, \sigma\rangle$ and every $\epsilon > 0$, there exists a deterministic persuasion rule f^* such that $m(f^*) < \inf_f m(f) + \epsilon$.

Proof. (i) By Lemma 1, there exists an optimal persuasion rule with a finite number of statements that induces acceptance with positive probability. Consider an optimal persuasion rule f with the fewest non-integer values. Let $0 < \alpha_1 < \cdots < \alpha_K < 1$ be the values of f that are not 0 or 1. We show that $K = 0$. If not, consider the set $T = \{s \mid f(s) = \alpha_1\}$. Let Y be the set of all states in which it is optimal for the speaker to make a statement from T, that is, $Y = \{x \mid \alpha(f, x) = \alpha_1\}$.

If the probability of $Y \cap A$ is at least that of $Y \cap R$, then consider f^+ which is a revision of f :

$$f^+(s) = \alpha_2 \text{ for all } s \in T \text{ and } f^+(s) = f(s) \text{ for } s \notin T.$$

Thus, $\alpha(f^+, x) = \alpha_2$ for $x \in Y$ and $\alpha(f^+, x) = \alpha(f, x)$ for $x \notin Y$. It follows that $m(f^+) \leq m(f)$.

If the probability of $Y \cap A$ is at most that of $Y \cap R$, then consider f^- which is a revision of f:

$$f^-(s) = 0 \text{ for all } s \in T \text{ and } f^-(s) = f(s) \text{ for } s \notin T.$$

Thus, $\alpha(f^-, x) = 0$ for $x \in Y$ and $\alpha(f^-, x) = \alpha(f, x)$ for $x \notin Y$. It follows that $m(f^-) \leq m(f)$.

The number of non-integer values used by either f^+ or f^- is reduced by 1, which contradicts the assumption that f uses the the minimal number of non-integer values.

(ii) Let f' be a persuasion rule such that $m(f') < \inf_f m(f) + \epsilon/2$. Let n be an integer such that $1/n < \epsilon/2$. Let f'' be the persuasion rule defined by $f''(s) = \max\{m/n \mid m/n \leq f'(s)\}$. Obviously $m(f'') < m(f') + 1/n$. The persuasion rule f'' involves a finite number of values. By the proof of Proposition 1 there is a deterministic persuasion rule f^* with $m(f^*) \leq m(f'')$. Thus, $m(f^*) < m(f') + \epsilon/2 < \inf_f m(f) + \epsilon$. □

Example 1 revisited: a solution. We return now to Example 1 and show that no persuasion rule induces a probability of error less than $\frac{4}{32}$. Consider an optimal persuasion rule that is deterministic. Thus, μ_x is either 0 or 1 for any state x. By the L-principle,

$\mu_{(1,1,1,0,0)} + \mu_{(1,1,0,0,0)} + \mu_{(1,0,1,0,0)} + \mu_{(0,1,1,0,0)} \geq 1$ and similar inequalities hold for any of the other 9 states in which exactly three aspects support the speaker. Summing up over these 10 inequalities yields

$$\sum_{n(x)=3} \mu_x + 3 \sum_{n(x)=2} \mu_x \geq 10.$$

Using the fact that μ_x is either 0 or 1 implies that $\sum_{n(x)=3} \mu_x + \sum_{n(x)=2} \mu_x \geq 4$ and thus $\sum_x p_x \mu_x \geq \frac{4}{32}$.

Let us now describe an optimal persuasion rule for this case. Partition the set of random variables into the two sets $\{1,2,3\}$ and $\{4,5\}$. The listener is persuaded only if the speaker can show that two random variables from the same cell of the partition support him. In states in which there are at least three random variables in favor of the speaker, at least two of them must belong to the same cell and, thus, the speaker is justifiably able to persuade the listener. However, in the four states in which exactly two random variables belonging to the same cell support the speaker's position, the speaker is able to persuade the listener even though he should not be able to. Thus, the probability of error under this persuasion rule is $\frac{4}{32}$.

This persuasion rule seems to be attractive when the partition of the random variables is prominent. For example, if the random variables are associated with Alice, Beth, Christina, Dan, and Edward, they can naturally be divided into two groups by gender. Given the constraint that the speaker cannot refer to more than two individuals, we have found an optimal persuasion rule whereby referring to two individuals of the same gender is more persuasive than referring to two individuals of different genders. ◊

EXAMPLE 3 (Persuading someone that the median is above the expected value). A speaker wishes to persuade the listener that the median of the values of three independent random variables uniformly distributed over the interval $[0,1]$ is above 0.5. The speaker can reveal the value of only one of the three random variables. Is it more persuasive to present a random variable with a realization of 0.9 or one with a realization of 0.6?

Formally, let $X = [0,1] \times [0,1] \times [0,1]$ with a uniform distribution and $A = \{(x_1, x_2, x_3) \mid$ two of the values are above 0.5$\}$. Let $x_i = t_i$ denote the statement "the realization of the variable x_i is t_i" and $S(t_1, t_2, t_3) = \{x_1 = t_1,\ x_2 = t_2,\ x_3 = t_3\}$. In other words $\langle X, S \rangle$ is vectoric.

The persuasion rule according to which the listener is persuaded only by the statement $x_1 = t_1$ where $t_1 > \frac{1}{2}$ yields a probability of error of $\frac{1}{4}$. We will employ the L-principle to show that this persuasion rule is optimal.

Note that the space X is isomorphic to the probabilistic space $Y \times Z$ with a uniform distribution, where $Y = [0, \frac{1}{2}] \times [0, \frac{1}{2}] \times [0, \frac{1}{2}]$ and $Z = \{-1,1\} \times \{-1,1\} \times \{-1,1\}$, by identifying a pair (y,z) with $x = (\frac{1}{2} + y_i z_i)_{i=1,2,3}$.

As a result, every $(y,(1,1,-1)) \in A$ is a part of an L with $(y,(-1,1,-1)) \in R$ and $(y,(1,-1,-1)) \in R$.

Thus we obtain the following inequalities:

$$\mu_{(y,(1,1,-1)} + \mu_{(y,(-1,1,-1))} + \mu_{(y,(1,-1,-1))} \geq 1$$
$$\mu_{(y,(1,-1,1)} + \mu_{(y,(-1,-1,1))} + \mu_{(y,(1,-1,-1))} \geq 1$$
$$\mu_{(y,(-1,1,1)} + \mu_{(y,(-1,1,-1))} + \mu_{(y,(-1,-1,1))} \geq 1.$$

Hence

$$\mu_{(y,(1,1,-1))} + \mu_{(y,(1,-1,1))} + \mu_{(y,(-1,1,1))} + 2\mu_{(y,(-1,1,-1))} + 2\mu_{(y,(1,-1,-1))} + 2\mu_{(y,(-1,-1,1))} \geq 3.$$

For deterministic persuasion rules it must be that at least two of the variables $\mu_{(y,z)}$ take the value 1 and, thus, for all y, we have $\sum_z p_{(y,z)}\mu_{(y,z)} \geq \frac{2}{8} = \frac{1}{4}$. If there exists a persuasion rule that yields an error probability strictly less than $\frac{1}{4}$, then by Proposition 1(ii) there is also a deterministic persuasion rule that yields an error probability less than $\frac{1}{4}$. Thus, the persuasion rule described above (which yields an error probability of exactly $\frac{1}{4}$) is optimal. ◊

5. A PROCEDURE FOR FINDING AN OPTIMAL PERSUASION RULE

We are now able to prove a proposition that reduces the task of finding an optimal persuasion rule to a simple optimization problem.

PROPOSITION 2. *Let $\langle X, A, p, \sigma\rangle$ be a finite persuasion problem. Let $(\mu_x^*)_{x\in X}$ be a solution to the optimization problem*

$$\min_{\{\mu_x\}_{x\in X}} \sum_{x\in X} p_x\mu_x \text{ s.t. } \mu_x \in \{0,1\} \text{ for all } x \in X \text{ and } \sum_{t\in\{x\}\cup T} \mu_t \geq 1 \text{ for any minimal } L,(x,T).$$

Then there is an optimal persuasion rule that induces the probabilities of errors $(\mu_x^*)_{x\in X}$.

PROOF. By Proposition 1 we can restrict ourselves to deterministic mechanisms. By Lemma 2 any persuasion rule satisfies the constraints (regarding the L's), so it is sufficient to construct a persuasion rule f that induces the optimal error probabilities vector $(\mu_x^*)_{x\in X}$.

Define $f(s) = 1$ for any signal s such that there exist $x \in A$ with $s \in \sigma(x)$ so that $\mu_x^* = 0$ and $\mu_y^* = 1$ for all $y \in R$ with $s \in \sigma(y)$. Define $f(s) = 0$ for any other signal s.

It is sufficient to show that for all x, the induced probability $\mu_x(f) \leq \mu_x^*$.

Let $x \in A$ and $\mu_x^* = 0$. There is a statement $s_x \in \sigma(x)$ so that $\mu_y^* = 1$ for all $y \in R$ such that $s_x \in \sigma(y)$. Otherwise, there is an $L, (x, T)$ such that $\sum_{t\in\{x\}\cup T} \mu_t^* = 0$. Thus $f(s_x) = 1$ and $\mu_x(f) = 0$

Let $x \in R$ and $\mu_x^* = 0$. Then there is no $s \in \sigma(x)$ such that $f(s) = 1$ and thus $\alpha(f, x) = 0$ and $\mu_x(f) = \mu_x^*$. □

6. Ex-post optimality

So far we have assumed that the listener is committed to a persuasion rule. In what follows, we address the question of whether the listener's optimal persuasion rule is one that he would indeed follow were he able to reconsider his commitment after the speaker has made his statement.

To motivate this analysis consider the following example.

Example 4. The listener wishes to choose a guest for a TV news program. He is looking for a person with strong views about the issues of the day. There is a potential candidate who the listener knows is one of four types: "hawk" (H), "dove" (D), a "pretender" (M) who can pretend to be either a hawk or a dove, or "ignorant" (I). The listener is not interested in the candidate's political views, but only in whether he has clear views one way or the other, i.e., if he is type H or D. The probabilities of the types are $p(H) = p(D) = 0.2$ and $p(M) = p(I) = 0.3$.

The listener can interview the candidate, after which he must decide whether or not to invite him onto the show. During the interview the listener plans to ask the speaker to make a statement regarding his views on current issues. Assume that apart from remaining silent (action 0), type H can make only the statement h; D can make only the statement d; and M can make either statement h or d. Type I can only remain silent. Thus, $\sigma(H) = \{h, 0\}$, $\sigma(D) = \{d, 0\}$, $\sigma(M) = \{h, d, 0\}$, and $\sigma(I) = \{0\}$.

A "naïve" approach to this problem is the following: Given the statement s, the listener excludes the types that cannot make the statement s and makes the optimal decision given the probabilities. For example, the message d excludes types I and H and therefore implies that the conditional probability that the speaker is of type D is 0.4. The listener thus rejects the speaker. This approach yields a probability of error of 0.4.

Suppose that the listener can commit to how he will respond to the speaker's statement. It is easy to see that, in this example, the listener can reduce the probability of error to 0.3. The best persuasion rule is to invite the speaker to the show if and only if he makes the statement d or h. (This avoids the possibility that I is invited to the show but leaves the possibility that, in addition to H and D, M might be invited.).

Assume now that the listener is released from his commitment once a statement has been made. If he believes that M's strategy is to utter d, then the listener, upon hearing the statement d, should attribute a higher probability to the possibility that he is facing M than to the possibility that he is facing D. Therefore, in this case he should not follow the optimal persuasion rule and should reject the speaker if he makes the statement d. If, however, the listener believes that M randomizes with equal probability between uttering d and h, then the listener, upon hearing the message d (h), should attribute the probability $\frac{4}{7}$ to the possibility that he is facing type D (H) and, thus, should not deviate from the optimal persuasion rule. ◊

Note that the ex-post optimality of the optimal persuasion rule in this example hinges on the knife-edge condition that the speaker of type M randomizes with equal probability between h and d. This observation hints at the possibility that a persuasion problem might exist in which the listener's optimal persuasion rule is not ex-post

optimal. However, as the analysis below demonstrates, this is never the case for finite persuasion problems .

For a given persuasion problem $\langle X, A, p, \sigma \rangle$, consider the corresponding extensive persuasion game $\Gamma(X, A, p, \sigma)$. First, nature chooses the state according to p; the speaker is then informed of the state x and makes a statement from the set $\sigma(x)$; and finally, after hearing the speaker's statement, the listener chooses between a and r. The payoff for the speaker is 1 if the listener takes the action a and 0 otherwise. The payoff for the listener is 1 if $x \in A$ and the action a is taken or if $x \in R$ and the action r is taken, and 0 otherwise. We say that a persuasion rule f is *credible* if there exists a sequential equilibrium of $\Gamma(X, A, p, \sigma)$ such that the listener's strategy is f.

EXAMPLE 2 REVISITED. The optimal persuasion rule described above is credible. The speaker's strategy of arguing in state (t_1, t_2) that $x_1 = t_1$ if $t_1 \geq t_2$ and that $x_2 = t_2$ if $t_2 > t_1$ is optimal. The set of types that use the argument $x_1 = t_1$ is $\{(t_1, x_2) \mid x_2 \leq t_1\}$. Conditional on this set, the probability that (t_1, x_2) is in A is greater than $\frac{1}{2}$ if and only if $t_1 > \frac{2}{3}$ and is less than $\frac{1}{2}$ if and only if $t_1 < \frac{2}{3}$. ◊

PROPOSITION 3. *If the persuasion problem is finite, then any optimal persuasion rule is credible.*

This proposition follows from solving the auxiliary problem presented in the next section.

COMMENT. The problem studied here can be viewed as a special case of a leader-follower problem in which the leader can commit to his future moves. As is well known, it is generally not true that the solution to such an optimization problem is credible. We are not aware, however, of any general theorem or principle that addresses this issue and that can explain why it is the case that in our model the listener's optimal strategy is credible. This question remains for future research.

We should emphasize, however, that Proposition 3 does not hold in case the listener has three actions, the speaker holds a fixed ordering over the actions, and the listener's preferences depend on the state. Consider, for example, the case in which the set of states is $X = \{1, 2\}$, the probability measure over X is $p_1 = 0.4$ and $p_2 = 0.6$, the signal function is $\sigma(1) = \{1\}$, $\sigma(2) = \{1, 2\}$, and the listener's set of actions is $\{1, 2, 3\}$. The speaker always prefers 1 over 2 and 2 over 3 and the listener's utility from the state x and action a is $u(1,1) = u(2,2) = 1$, $u(1,2) = u(2,1) = -1$, and $u(1,3) = u(2,3) = 0$. The optimal persuasion rule for the listener is to respond to signal 2 with action 2 and to signal 1 with action 3. However, once he observes signal 1 it is better for the listener to take action 1.

7. THE BRIDGES PROBLEM

A group of individuals is partitioned into a finite number of types, which are members of a set X. The mass of type x is p_x. Let S be a set of bridges spanning a river. The individuals are located on one side of the river and would like to cross to the other side.

Individuals of type $x \in X$ can use only the bridges in the set $\sigma(x) \neq \emptyset$. The set X is partitioned into two subsets, A whose members are welcome on the other side and R whose members are not. A decision maker has to decide, for each bridge, the probability that that bridge will be open. The decision maker cannot discriminate between the individuals in A and R. Each individual of type x chooses a bridge in $\sigma(x)$ with the highest probability of being open from among the ones he can use. The decision maker's objective is to maximize the "net flow", i.e., the difference in size between the group of type A's and the group of type R's crossing the river.

A bridge policy determines the probability with which each bridge is open. A bridge policy is credible if there exists an assignment of types to bridges whereby: (*i*) each type is assigned only to a bridge he can use, (*ii*) within the set of bridges he can use, each type is assigned only to bridges with the highest probability of being open, and (*iii*) the mass of types in A who are assigned to a bridge that is open (closed) with strictly positive probability is at least as high (low) as the mass of types in R who are assigned to that bridge. We show that any optimal bridge policy is credible.

Formally, a *bridge policy* is a function $O : S \to [0,1]$ with the interpretation that $O(s)$ is the probability that bridge s is open. Let $\alpha(O,x) = \max\{O(s) \mid s \in \sigma(x)\}$, that is the maximal probability of crossing the bridges that type x can achieve given the bridge policy O. Let $N(O) = \sum_{x\in A} p_x \alpha(O,x) - \sum_{x\in R} p_x \alpha(O,x)$ be called the net flow. A bridge policy is *optimal* if it maximizes $N(O)$. Given a bridge policy O, a *rational feasible bridge assignment* β is a function that assigns to each type x a probability measure on $\sigma(x)$, such that $\beta(x)(s) > 0$ only for values of s that maximize $O(s)$ in $\sigma(x)$. Given an assignment β, the *net assignment* to bridge s is $n(s,\beta) = \Sigma_{x\in A} p_x \beta(x)(s) - \Sigma_{x\in R} p_x \beta(x)(s)$. A bridge policy O is *credible* if there is a rational feasible assignment β such that for every s, $O(s) > 0$ implies $n(s,\beta) \geq 0$ and $O(s) < 1$ implies $n(s,\beta) \leq 0$.

CLAIM 1. *All optimal bridge policies are credible.*

PROOF. Let O^* be an optimal bridge policy. For any assignment β, let

$$\delta(\beta) = \sum_{s\in\{s \mid n(s,\beta)<0\}} |n(s,\beta)| O^*(s) + \sum_{s\in\{s \mid n(s,\beta)>0\}} n(s,\beta)(1-O^*(s)).$$

Let β^* be a minimizer of $\delta(\beta)$ over all rational feasible assignments. We show that $\delta(\beta^*) = 0$ and thus for all s such that $O^*(s) > 0$ we have $n(s,\beta) \geq 0$ and for all such s that $O^*(s) < 1$ we have $n(s,\beta) \leq 0$.

Assume, for the purpose of contradiction, that $\delta(\beta^*) > 0$. Assume that there is a bridge s for which $O^*(s) > 0$ and $n(s,\beta^*) < 0$ (an analogous argument applies to the case in which there is a bridge s for which $O^*(s) < 1$ and $n(s,\beta^*) > 0$).

Let α be the minimum of $O^*(s)$ over $\{s \mid O^*(s) > 0 \text{ and } n(s,\beta^*) < 0\}$. Let $S(\alpha) = \{s \mid O^*(s) = \alpha\}$. Let $X(\alpha) = \{x \mid \beta^*(x)(s) > 0 \text{ for a bridge } s \text{ such that } s \in S(\alpha)\}$, that is, $X(\alpha)$ is the set of types who are assigned by β^* to the bridges whose probability of being open is α. Note that types in $X(\alpha)$ cannot do better than trying to cross a bridge in $S(\alpha)$ and are indifferent between all bridges in $S(\alpha)$. Let $S_0 = \{s \in S(\alpha) \mid n(s,\beta^*) < 0\}$. The set S_0 is not

empty and contains all bridges that are open with probability α and for which the net assignment is negative.

Let $y_1,\dots,y_T$ be the longest sequence of distinct bridges in $S(\alpha)-S_0$ such that for every y_t,

(i) $n(y_t,\beta^*)=0$

(ii) there exist $x\in R$ and $y_0\in S_0\cup\{y_1,\dots,y_{t-1}\}$ such that $\beta^*(x)(y_0)>0$ and $y_t\in\sigma(x)$.

In other words, under β^* each y_t is a bridge with a zero net transfer such that there is a positive mass of types in R that can cross y_t and is assigned by β^* either to cross a bridge that precedes y_t in the sequence or to cross a bridge in S_0.

Denote $Z=S_0\cup\{y_1,\dots,y_T\}$. There are two possibilities:

(i) There is no $s\in S(\alpha)-Z$, $x\in R$, and $z\in Z$ such that $s\in\sigma(x)$ and $\beta^*(x)(z)>0$. That is, there is no bridge s outside Z that is opened with probability α and that can be crossed by a type in R who can cross the river with probability α. The net transfer in Z is negative. Reducing the probability of transfer to all bridges in Z will increase the total net flow, thus violating the optimally of O^*.

(ii) There is $s\in S(\alpha)-Z$, $x\in R$, and $z\in Z$ such that $s\in\sigma(x)$ and $\beta^*(x)(z)>0$. By the definition of $(y_1,\dots,y_T)$ it must be that $n(s,\beta^*)>0$. It follows that there are sequences of distinct bridges $s_0,s_1,\dots,s_K=s$ and types $i_0,\dots,i_{K-1}\in R$ such that $s_0\in S_0$, $\beta^*(i_k)(s_k)>0$, and $s_{k+1}\in\sigma(i_k)$ (for $k=0,\dots,K-1$). This allows us to construct a new rational assignment β by shifting a positive mass of types in R from s_0 to s_1, from s_1 to s_2, and so on, such that $\delta(\beta)<\delta(\beta^*)$. Formally, let ε be a positive number such that for $k=0,\dots,K-1$ we have $\varepsilon<\beta^*(i_k)(s_k)$, $\varepsilon<n(s_K,\beta^*)$, and $\varepsilon<|n(s_0,\beta^*)|$. Define β as an assignment that is obtained from β^* by successively shifting to s_{k+1} a mass ε of individuals of type i_k assigned by β^* to cross s_k. For all bridges with the exception of s_0 and s_K we have $n(s,\beta)=n(s,\beta^*)$. Furthermore, $n(s_K,\beta)=n(s_K,\beta^*)-\varepsilon>0$ and $n(s_0,\beta)=n(s_0,\beta^*)+\varepsilon<0$. Thus, $\delta(\beta)=\delta(\beta^*)-\alpha\varepsilon-(1-\alpha)\varepsilon$, contradicting the choice of β^*.

Thus, it follows that there exists a rational feasible assignment with nonnegative net flow on all open bridges and nonpositive net flow on all closed bridges. □

8. Concluding remarks

This paper has attempted to make a modest contribution to the growing literature linking economic theory to linguistics. Our purpose is not to suggest a general theory for the pragmatics of persuasion but rather to demonstrate a rationale for inferences in persuasion situations.

One of our main findings is that any optimal persuasion rule is also ex-post optimal. It is quite rare that in a principal-agent problem the optimal incentive scheme is one that the principal would wish to obey even after the agent has made his move. The bridge problem described in Section 7 provides an example of a principal-agent problem that

in fact does have this property. The problem discussed in Glazer and Rubinstein (2004) is shown there to have this property as well. The generalizability of this result is still an open question.

Our work is related to several areas of research in linguistics and economics. In the linguistics literature, our paper belongs to the emerging field that tries to explain pragmatic rules by employing game theoretical methods. In our approach, pragmatic rules determine a game between the participants in the discourse. Whatever the process that created these rules, it is of interest to compare them with the rules that would have been chosen by a rational designer seeking to maximize the functionality of the discourse. Such an approach is suggested in Glazer and Rubinstein (2001, 2004) and discussed in Rubinstein (2000). A recent collection of articles in Benz et al. (2006) presents various ideas that explain pragmatics phenomena using game theoretical tools.

Within the economic literature our paper is related to two areas of research.

The first investigates sender-receiver games (see Crawford and Sobel 1982) in which one agent (the sender) sends a costless message to the other (the receiver). The receiver cannot verify any of the information sent by the sender and the interests of the sender and the receiver do not necessarily coincide. The typical question in this literature is whether an informative sequential equilibrium exists.

The second (and closer) area of research studies models where a principal tries to elicit verifiable information from the agent(s). The agent however can choose which pieces of information to convey. Among the early papers on this topic are Townsend (1979), Green and Laffont (1986), and Milgrom and Roberts (1986), and among the more recent are Bull and Watson (2004), Deneckere and Severinov (2003), Fishman and Hagerty (1990), Forges and Koessler (2005), Lipman and Seppi (1995), and Shin (1994).

References

Benz, Anton, Gerhard Jäger, and Robert van Rooij, eds. (2006), *Game Theory and Pragmatics*. Palgrave Macmillan, Basingstoke. [409]

Bull, Jesse and Joel Watson (2004), "Evidence disclosure and verifiability." *Journal of Economic Theory*, 118, 1–31. [409]

Crawford, Vincent P. and Joel Sobel (1982), "Strategic information transmission." *Econometrica*, 50, 1431–1451. [409]

Deneckere, Raymond and Sergei Severinov (2003), "Mechanism design and communication costs." Working paper, Fuqua School of Business, Duke University. [409]

Fishman, Michael J. and Kathleen M. Hagerty (1990), "The optimal amount of discretion to allow in disclosure." *Quarterly Journal of Economics*, 105, 427–444. [409]

Forges, Françoise and Frédéric Koessler (2005), "Communication equilibria with partially verifiable types." *Journal of Mathematical Economics*, 41, 793–811. [409]

Glazer, Jacob and Ariel Rubinstein (2001), "Debates and decisions: On a rationale of argumentation rules." *Games and Economic Behavior*, 36, 158–173. [396, 409]

Glazer, Jacob and Ariel Rubinstein (2004), "On optimal rules of persuasion." *Econometrica*, 72, 1715–1736. [396, 398, 400, 401, 402, 409]

Green, Jerry R. and Jean-Jacques Laffont (1986), "Partially verifiable information and mechanism design." *Review of Economic Studies*, 53, 447–456. [409]

Grice, H. Paul (1989), *Studies in the Way of Words*. Harvard University Press, Cambridge, Mass. [395]

Lipman, Barton L. and Duane J. Seppi (1995), "Robust inference in communication games with partial provability." *Journal of Economic Theory*, 66, 370–405. [409]

Milgrom, Paul and John Roberts (1986), "Relying on the information of interested parties." *Rand Journal of Economics*, 17, 18–32. [409]

Rubinstein, Ariel (2000), *Economics and Language*. Cambridge University Press, Cambridge. [409]

Shin, Hyun Song (1994), "The burden of proof in a game of persuasion." *Journal of Economic Theory*, 64, 253–264. [409]

Townsend, Robert M. (1979), "Optimal contracts and competitive markets with costly state verification." *Journal of Economic Theory*, 21, 265–293. [409]

Submitted 2006-3-3. Final version accepted 2006-7-3. Available online 2006-7-4.