

Classification of natural language messages using a coordination game

Daniel Houser · Erte Xiao

Received: 25 November 2007 / Accepted: 1 September 2010
© Economic Science Association 2010

Abstract The role of natural language communication in economic exchange has been the focus of substantial experimental analysis. Recently, scholars have taken the important step of investigating whether certain types of communication (e.g., promises) might affect decisions differently than other types of communication. This requires classifying natural language messages. Unfortunately, no broadly-accepted method is available for this purpose. We here describe a coordination game for classification of natural language messages. The game is similar in spirit to the “ESP” game that has proven successful for the classification of tens of millions of internet images. We compare our approach to self-classification as well as to classifications based on a standard content analysis. We argue that our classification game has advantages over those alternative approaches, and that these advantages might stem from the salient rewards earned by our game’s participants.

Keywords Classification game · Natural language · Messages · Coordination game · Experimental methods

1 Introduction

The role of natural language communication in economic exchange has been the focus of substantial experimental analysis. By exogenously varying the possibility

Electronic supplementary material The online version of this article (doi:[10.1007/s10683-010-9254-4](https://doi.org/10.1007/s10683-010-9254-4)) contains supplementary material, which is available to authorized users.

D. Houser (✉)
George Mason University, Fairfax, USA
e-mail: dhouser@gmu.edu

E. Xiao
Carnegie Mellon University, Pittsburgh, USA
e-mail: exiao@andrew.cmu.edu

of communication, an early and continuing experimental literature provides rigorous data on its importance to economic outcomes (see, e.g., Dawes et al. 1977; Issac and Walker 1988; Ben-Ner et al. 2007; Cason and Mui 2007, 2010). More recently, economists have taken the important step of investigating whether certain types of communication (e.g., promises) might affect decisions differently than other types of communication (e.g., empty talk). This requires classifying natural language communication. Unfortunately, no broadly-accepted method is available for this purpose. We here describe a procedure for the classification of natural language messages using a coordination game. We illustrate this procedure using data reported by Charness and Dufwenberg (2006). Further, we compare our coordination-game message classification procedure with two alternative methods commonly used in communication research: content analysis and researchers' self-classification. We provide evidence that classification using a coordination game allows one to learn more from the data than is possible with these alternatives.

Economic theory on communication as an abstract signal emerged early (e.g., von Neumann and Morgenstern 1944, especially Sect. 6.4.3; Luce and Raiffa 1957, Chap. 6). Pioneering experimental research by Fouraker and Sigel (1963, Chap. 4) provided early evidence that even restricted forms of communication could have substantial impact on behavior (see also Cason and Mui 2007). Economic experiments allowing subjects to communicate using natural language appeared shortly thereafter (see e.g., Ledyard 1995, Sect. III.3, and Holt 1995, Sect. VIII.D), and convincingly demonstrated that allowing communication can foster cooperation (or collusion) in a variety of environments. Ledyard (1995, p. 158) points out that the question left unanswered by this research is why.

Very recently, a number of scholars have investigated reasons for communication's effects (e.g., Xiao and Houser 2005, 2007; Charness and Dufwenberg 2006; Kimbrough et al. 2008; Schotter and Sopher 2007; Ellingsen and Johannesson 2007; Brandts and Cooper 2007; Sutter and Strassmair 2009; Hennig-Schmidt et al. 2008; Cooper and Kagel 2005). Each of these papers follows the approach of classifying natural language messages written during experiments in ways that substantively inform the authors' scientific hypotheses. The classification procedures they use differ.

For example, Charness and Dufwenberg (2006, discussed further below), Schotter and Sopher (2007) and Kimbrough et al. (2008) all self-classify their natural language messages. These classification data were then used to draw, respectively, the following conclusions: promises increase economic efficiency, fairness is a justification for rejections in ultimatum games, and people distinguish between personal and impersonal exchange environments.

Others have used "content analysis," an approach pioneered by psychologists and sociologists (see, e.g., Krippendorff 2004). In a typical content analysis, two or three research assistants are recruited and trained to code the messages using categories developed by the researchers. The coders usually work independently. For example, Cooper and Kagel (2005) studied team versus individual play in signaling games. The authors developed categories, and then recruited two research assistants to classify messages according to those categories. Similar classification methods are used in Brandts and Cooper (2007), Sutter and Strassmair (2009), and Hennig-Schmidt et al. (2008).

We argue here that our classification coordination game can, at some level of cost, clarify and extend the nature of conclusions reached in these sorts of analyses. For example, in the Charness and Dufwenberg (2006) data we study below, our classification coordination game forced us to articulate a clear answer to “what is a promise?” It turns out that doing this enables us to extend, in a useful way, Charness and Dufwenberg’s original findings.

Furthermore, in comparison with traditional content analysis procedures that pay a flat rate for coding, we find coders are more sensitive to instructions under the XH coordination game. As a result, using the coordination game seems relatively more powerful in distinguishing subtleties in the meanings of written messages.

2 Classification coordination game (XH classification game)

Xiao and Houser (2005, henceforth XH) suggest a laboratory coordination game for the classification of natural language messages. To the best of our knowledge, ours is the first such suggestion to be made within the context of message classification. The game, however, is related to the well-known “ESP” game (von Ahn 2005), which has been licensed by Google and is functioning successfully as their Google Image Labeler.¹ The XH game has been used for message classification by Xiao and Houser (2005, 2007), as well as Ellingsen and Johannesson (2007), Corazzini et al. (2009), and Engle-Warnick and Soroka (2009), among others. An advantage of the XH game is that incentives motivate coders to classify messages carefully.

The XH classification game proceeds as follows. A group of $N > 2$ evaluators is given a list of $M > 0$ messages and a set of $K > 1$ categories. The goal is to assign each message to a single category. Evaluators are told that after they have all completed this task, a set of $J > 0$ messages will be chosen at random. Each evaluator earns $\$Y > 0$ for each of the chosen messages, such that their own message classification matches a most popular classification.

The XH game is a coordination game (for surveys see Ochs 1995; Crawford 1997; Camerer 2003). In particular, there are a large number of Nash equilibria varying according to the beliefs people bring to the experiment. For example, if “Category 1” becomes a focal point, so that all evaluators believe most evaluators will assign all messages to “Category 1,” then that strategy is an equilibrium of this game.

¹In the ESP game, participants are randomly paired. Matched participants are anonymous and cannot communicate. Each member of the pair is shown the same image. They both begin to enter words as candidate labels for the image. A word that is entered by both (not necessarily at the same time) becomes the image’s label (or part of the label). The task is to label as many images as possible within a fixed amount of time. Thus, the ESP game is a coordination game with multiple equilibria. For example, if both people in the pair believe both people in the pair will use the first word on page 124 of Epstein (2007) as an image’s label, then using the word “uniform” as the label is the equilibrium. This is, in our view, unlikely. More likely is that people will coordinate on the content of the image, and suggest labels that are sensible with respect to expected community standards and opinions. In fact, the now tens of millions of successful labels indicate that this is exactly what happens. It is also what happens in our classification game. Note that Feiler and Camerer (2009) use a similar paradigm to investigate how subjects from two different campuses coordinate on naming a mixture of pictures from both campuses.

Like the ESP game, the interesting equilibrium of the XH game emerges when all evaluators believe that all evaluators will classify the natural language messages according to the content of the message. Under these beliefs, evaluators must read and consider each message and classify it in a way that they think is sensible by “community” standards.² For most cases of practical importance, including the issues concerning all of the papers discussed above, this is the relevant standard when drawing an inference regarding the meaning of a natural language message.

Given that there is a single equilibrium of interest, but many that exist, it is natural to ask how people actually play this game. This is an empirical question, and the empirical evidence gathered to date, including the data in analysis described in Sect. 3 below, make a cogent argument that evaluators coordinate on message content.

An alternative objective procedure to message classification is to define word-based rules that determine the category into which a message falls. For example, in the context of our analysis below, one could specify that a message is to be classified as a “Promise” if the message contains the word “promise,” but only if “promise” is not preceded by “not.”³ The XH game provides an approach to message classification that is especially appealing when scholars are either unwilling or unable to specify rules linking the words that comprise a message to that message’s meaning.

The next section illustrates the XH game using data on natural language promises reported by Charness and Dufwenberg (2006). We also conduct a more traditional content analysis of the same set of messages. With both procedures, evaluators were divided into two groups with instructions that differed in terms of what sort of message should be classified as a promise. In doing this, we demonstrate that our classification method can shed light on hypotheses that self-classification cannot inform. Moreover, in relation to traditional content analysis, we find that classifications stemming from the XH game are more sensitive to differences in the classification criteria, and in a way that has substantive implications for the nature of the results.

3 Illustrating the XH coordination game

To illustrate the XH coordination game we draw from a recent contribution by Charness and Dufwenberg (2006) (henceforth C&D) which reports data from laboratory experiments to test the impact of communication in a one-shot principal-agent game. C&D find that messages sent from principals to agents affect beliefs, and that different beliefs lead to different actions (pp. 1587–1589). Moreover, they find that beliefs change when messages contain promises or statements of intent (pp. 1590–1591). C&D conclude that people are sensitive to guilt in their experiment, so that promises sent from principals to agents enhance trust, cooperation and efficiency (p. 1594).

²Google’s goal with the ESP game is to make image searches faster and more accurate by associating with each image a set of key-words that are sensible by “community” standards. Google’s continued investment in this project suggests they are optimistic that this works.

³Efforts to classify by text-parsing were used early-on in the empirical reputation literature, where people attempted to gather vast amounts of auction data (usually from eBay) using ‘spydery’ and then categorize those data using automated text-parsing routines. Doing this turns out to be extremely difficult and unreliable. Houser and Wooders (2006) suggest an alternative approach.

Table 1 Payoff table for C&D “(5, 5)” game

	A: <i>In</i>		B: <i>Roll</i>	
	A: <i>Out</i>	B: <i>Don't Roll</i>	<i>Failure</i> [$p = 1/6$]	<i>Success</i> [$p = 5/6$]
Payoff (A, B)	(5, 5)	(0, 14)	(0, 10)	(12, 10)

In C&D’s analysis whether a given message includes a promise or statement of intent is determined by C&D’s own evaluation. Below, we use both the XH message classification game, as well as a more traditional content analysis, to re-analyze data from one of the several treatments reported by C&D. We narrow our focus further by restricting our attention to inferences regarding exchange efficiency, only one of several dimensions C&D report in their paper. The analysis is sufficient for our purpose of demonstrating the XH classification game.

3.1 C&D’s game

C&D study one-shot sequential two-player games using the strategy method. Player “A” chooses *In* or *Out*, and player “B” chooses to either *Roll* or *Don't Roll* a six-sided die. Player B’s choice affects payoffs only if “A” chooses *In*. Player “B” makes her decision without knowing player A’s actual choice, but under the hypothetical condition that “A” chose *In*. Table 1 details the payoffs in C&D’s “(5, 5)” treatment.

The unique efficient outcome of the game occurs when “A” chooses “*In*” and “B” chooses “*Roll*.” C&D investigate behavior in treatments where player “B” can send cheap-talk pre-play messages to player “A.” C&D self-classify messages as including a promise or not, and after doing so find that “B” is more likely to *Roll*, and “A” is more likely to choose *In*, when “B” sends a pre-play promise to *Roll*. Thus, C&D conclude that promises enhance economic efficiency.

We report below two analyses of C&D’s message data using the XH classification game. In one analysis, evaluators are instructed to classify a message as a promise if it includes “any statement of intent” (we denote this as the Weak Promise treatment). We find that messages classified as “promises” in this condition increase economic efficiency, though the effect is not statistically significant.

We also obtain classifications of the same messages using the XH classification game, but under a treatment where evaluators are instructed to classify a message as a promise only if that message is a clear promise (we call this the Strong Promise treatment). The resulting classifications are closely aligned with C&D, and we obtain the result that promises statistically significantly enhance economic efficiency.

Finally, we conduct a content analysis of those same messages using the same two criteria (strong promise vs. weak promise). We find that the resulting classifications do not vary with differences in the classification criteria.

3.2 Coordination game message classification procedures

We recruited a total of 50 evaluators from the general student population at George Mason University, 25 each for the Weak and Strong Promise treatments. Evaluators

were seated at visually separated desks where they worked independently. Before receiving any messages, subjects were acquainted with the C&D game and were provided with a transcript of C&D's instructions for their (5, 5) "Messages" treatment. We used a quiz to ensure everyone understood the C&D game.

Messages were transcribed directly from the C&D supplementary material on the *Econometrica* website, and were presented to evaluators in different and random orders. Evaluators were not given any information regarding the decisions of the message-writers or their counterparts. Also, evaluators were not given any information regarding the purpose of the study, the hypotheses of interest or C&D's previous evaluations. Evaluators were instructed to classify each message as either "Promise or Intent" or "Empty Talk" (subjects' instructions are reproduced in the [Electronic Supplementary Material \(ESM\)](#).)

Our "Weak Promise" treatment attempts to follow closely the approach described by C&D (see page 1590 of their article). In particular, evaluators were verbally instructed: "You should classify a message as "Promise or Intent" if, in your opinion, it includes any statement of intent." In our "Strong Promise" treatment, we used the same written instructions, but in this case the experimenter told evaluators, "You should classify a message as "Promise or Intent" only if, in your opinion, it is certainly a promise."

Messages were classified using the XH message classification game. Each evaluator was instructed to evaluate every message, and was told she would earn \$5 for doing so. Evaluators knew that three messages would be randomly chosen at the end of each session. For each, if an evaluator's classification matched the most popular classification, she earned an additional \$5 (so up to \$15 could be earned this way). The evaluation procedure took about an hour and median salient earnings were \$15, or \$22 total including the \$7 all subjects earned for arriving to the lab on time.

3.3 Content analysis procedures

We recruited another 49 evaluators⁴ from the same subject pool at George Mason University to participate in the content analysis of the message (24 for the strong promise treatment and 25 for the weak promise treatment). The procedures were similar to the XH game treatments, with the important exception that evaluators were given more detailed written instructions (see [Electronic Supplementary Material \(ESM\)](#) for details).⁵ The written instructions for the content analyses differed between the Strong and Weak Promise treatments. These differences were also emphasized verbally to the subjects using the same verbal instructions we used in the XH game treatments.

⁴Researchers often recruit two or three evaluators when applying content analysis (Krippendorff 2004). In our case, in order to compare the results of content analysis to our method, we kept the number of evaluators similar between treatments.

⁵Specific instructions are common in content analyses (Krippendorff 2004). We thank Tim Cason for thoughts and references regarding how one might pursue a content analysis in our environment. We chose to adopt a "minimal training" procedure in order to maintain maximal symmetry with our coordination game procedures (where there is no training). Presumably, it would be possible to "train" coders to provide responses that closely reflect any predetermined pattern, including that which arises from our coordination game.

Content analysis differs from our coordination game in multiple ways (e.g., evaluating according to what one thinks others believe as compared to what oneself believes). Note that one important difference in relation to the XH game is that evaluators' payoffs did not depend on others' evaluations of the messages. In particular, subjects were paid a fixed \$12 per hour of time spent classifying messages up to a maximum of two hours, and this constraint was not binding for any of our coders. Including the \$7 show-up bonus, total earnings in this experiment were \$19.

3.4 Results

3.4.1 *Coordination on content of message*

We noted above that XH is a coordination game, and that the equilibrium of interest occurs when evaluators coordinate on the content of each message. Our evidence is that this is indeed what occurs. In particular, in both treatments, when the message was a clear promise (e.g., it includes the word "promise"), evaluators labeled it a promise. On the other hand, when the message was clearly empty talk (e.g., entirely unrelated to the game) evaluators labeled it accordingly. In less clear cases, opinions broke toward "Promise" in the "Weak Promise" treatment, and toward "Empty Talk" when the requirement was stronger. One would predict this to occur when people coordinate using the content of the message.⁶ We are unable to discover any explanation for the patterns in our classification data other than coordination on content.

3.4.2 *Strong and weak promise classifications with XH and content analysis*

C&D obtained 42 observations in their (5, 5) treatment. However, four of these included no message. The 38 written messages were classified by our evaluators. We first examine how classifications change when the classification criterion changes (in our case, Weak Promise vs. Strong promise). We hypothesize that, in comparison with content analysis, coders in the XH classification game are more likely to respond to differences in the criteria. The reason is that in the XH game the coder's reward varies according to her decisions, while with content analysis the reward is fixed. Thus, we hypothesize that the XH game is advantaged by the presence of salient rewards.⁷

To test the hypothesis, we begin by classifying a message as a "Promise" (either strong or weak) when more than 50% of evaluators (in a particular treatment) code it as such.⁸ So classified, we find that with content analysis only one of 38 messages (2.6%) changed classification from "Promise" in the "Weak Promise" treatment to

⁶Evaluators also completed a questionnaire while waiting to be paid. Their responses are additional evidence that evaluations were based on the messages' content. For example, one evaluator wrote: "There were a few messages that made it difficult to determine the intent of player B. Most messages were pretty straightforward."

⁷This is a very familiar argument to experimental economists; Smith (1962) was first to emphasize the importance of salient rewards.

⁸One message received a tie in the content analysis in the strong promise treatment. We coded this message as "Empty Talk." Coding it as a "Promise" strengthens our conclusions regarding the differences between our method and content analysis.

“Empty Talk” in the “Strong Promise” treatment. In contrast, using the XH classification game, seven messages (18.4%) switched from “Promise” in the “Weak Promise” treatment to “Empty Talk” in the “Strong Promise” treatment (and note that the message that switched in Content Analysis also switched in XH). The difference in the frequency of switches is statistically significant (Z -test, $p = 0.02$).

One can also calculate, for each message, the percentage of coders who classified the message as “Promise” in the “Weak” and “Strong” Promise treatments. Doing this reveals that, on average, the frequency of “Promise” classifications is 18% higher in the “Weak” treatment than in the “Strong” treatment in the XH classification game, but only 2% higher in the “Weak” treatment under content analysis. This difference (between 18% and 2%) is statistically significant (Mann-Whitney test, $p = 0.00$, treating each message as an independent observation), supporting our hypothesis that coders in the XH game are more responsive to the classification criterion than with content analysis.

Finally, under content analysis we observe 14 of 38 messages (37%) where more evaluators classified the message as a promise in the “Strong Promise” treatment than in the “Weak Promise” treatment. This is unexpected, as the former category should be a subset of the latter. In contrast, only 2 of 38 messages (5.2%) under XH classifications have this property.

3.4.3 *Strong promises promote economic efficiency*

Among the 38 messages classified by our evaluators, C&D labeled 24 as “Promise” and the remaining 14 as “Empty Talk.” Our “Strong Promise” treatment using the XH game yielded identical frequencies. On the other hand, we obtained 31 “Promise” and 7 “Empty Talk” classifications in our “Weak Promise” treatment. Under content analysis, the frequencies of “promise” classifications are 26 and 27 in the “Weak” and “Strong” treatments, respectively (note the reversal: there are fewer “Weak” promises than “Strong” promises under content analysis). For the purpose of the statistical analyses below, we follow C&D and group the no-message cases with those messages classified as “Empty Talk.”

In both treatments, under either content analysis or with the XH game, most classifications are the same as used by C&D. Table 2 lists all the cases where classifications switched in relation to C&D’s. Using the XH game, in the “Weak Promise” treatment, all switches were from C&D’s label of “Empty Talk” to our evaluators’ labels of “Promise.” In the “Strong Promise” case, in relation to C&D’s classifications, there were two switches to “Empty Talk” and two switches to “Promise.”

With content analysis, in the “strong promise” treatment, three messages switched to “Promise” and one message switched to “Empty Talk” (in relation to C&D’s classifications). Each of these switches also occurred in the “Weak promise” treatment. In addition, one additional message switched to “Promise.”

C&D report that the frequency of economically efficient outcomes increases from 27% to 67% between cases where the message did not and did include a promise, and that this increase is statistically significant ($Z = 2.5$, $p < 0.01$, one-tailed test). Using their same procedures (a standard proportions test) and applying the classifications based on content analysis, we find that the magnitude of the promise effect

Table 2 Messages where C&D and evaluators have different classifications

Message	Classification		“P” frequency (%)
	C&D	Evaluators	
XH game			
• Weak Promise Treatment			
– Please choose In so we can get paid more.	E	P	15(60%)
– If you stay in, the chances of the die coming up other than 1 are 5 in 6—pretty good. Otherwise, we’d both be stuck at \$5. (If you opt out)	E	P	21(84%)
– If you choose IN the first round and then I will choose Don’t Roll at first. I will get \$14 but then after that I will choose roll each time after the 1st role. Chances are most likely you will get \$12 and I will get only \$10.I will the only take 7 rolls for you to get even with me. That way we both leave with a good amount of money. Hope you have a great evening and that this works out for both of us. ☺	E	P	21(84%)
– If you choose IN, and I roll, the chances of our getting the most \$ are very high. The likelihood of my rolling a 1 is small compared to the chances of rolling a 2–6. So we both get cash.	E	P	19(76%)
– Hopefully I’ll make a lucky role.	E	P	15(60%)
– Hello fair stranger, anonymous partner . . . Choose whatever you want. Far be it from me to influence your decision, but I think you should choose “in” and I should choose “roll” and we should take the chance at both earning as much as we can. 5 chances out of 6 say it’ll work, and I’m totally broke, looking to rake in stray cash however I can. I feel the luck in the air. I don’t really have much else to say. Hope you’re doing well, whoever you are. Yes. That’s all. Random note from random human	E	P	14(56%)
– If you choose <i>IN</i> you have the best opportunity to make the most money. You have a 5/7 chance of making more money! So <i>IN</i> would be your best bet. Cheers. ☺	E	P	16(64%)
• Strong Promise Treatment			
– If I roll a 2–6 (you’ll know when you receive the \$, you will give \$5.00 to a stranger. [[[then there is a line, under which is written “Sign here if you are so kind”]]]] Thanks. You’ll still be gaining more than if I had chosen Don’t roll.	P	E	6(24%)
– If you choose IN the first round and then I will choose Don’t Roll at first. I will get \$14 but then after that I will choose roll each time after the 1st role. Chances are most likely you will get \$12 and I will get only \$10. I will the only take 7 rolls for you to get even with me. That way we both leave with a good amount of money. Hope you have a great evening and that this works out for both of us.☺	E	P	19 (76%)
– Choose “In” so we can both make some \$\$ What are the chances me rolling a 1? I’ll try my best.	P	E	10(40%)
– If you choose IN, and I roll, the chances of our getting the most \$ are very high. The likelihood of my rolling a 1 is small compared to the chances of rolling a 2–6. So we both get cash.	E	P	14(56%)

Table 2 (continued)**Content Analysis****• Weak Promise Treatment**

– If you stay in, the chances of the die coming up other than 1 are 5 in 6—pretty good. Otherwise, we’d both be stuck at \$5. (If you opt out)	E	P	18(72%)
– If I roll a 2–6 (you’ll know when you receive the \$, you will give \$5.00 to a stranger.	P	E	9(36%)

[[[then there is a line, under which is written “Sign here if you are so kind”]]]

Thanks.

You’ll still be gaining more than if I had chosen Don’t roll.

– If you choose IN the first round and then I will choose Don’t Roll at first. I will get \$14 but then after that I will choose roll each time after the 1st role. Chances are most likely you will get \$12 and I will get only \$10.I will the only take 7 rolls for you to get even with me. That way we both leave with a good amount of money. Hope you have a great evening and that this works out for both of us. ☺	E	P	19(76%)
--	---	---	---------

– If you choose IN, and I roll, the chances of our getting the most \$ are very high. The likelihood of my rolling a 1 is small compared to the chances of rolling a 2–6. So we both get cash.	E	P	19(76%)
--	---	---	---------

Hello fair stranger, anonymous partner . . . Choose whatever you want. Far be it from me to influence your decision, but I think you should choose “in” and I should choose “roll” and we should take the chance at both earning as much as we can. 5 chances out of 6 say it’ll work, and I’m totally broke, looking to rake in stray cash however I can. I feel the luck in the air.	E	P	13(52%)
--	---	---	---------

I don’t really have much else to say. Hope you’re doing well, whoever you are.

Yes.

That’s all. Random note from random human

• Strong Promise Treatment

– If you stay in, the chances of the die coming up other than 1 are 5 in 6 – pretty good. Otherwise, we’d both be stuck at \$5. (If you opt out)	E	P	17(71%)
– If I roll a 2–6 (you’ll know when you receive the \$, you will give \$ 5.00 to a stranger.	P	E	8(33%)

[[[then there is a line, under which is written “Sign here if you are so kind”]]]

Thanks.

You’ll still be gaining more than if I had chosen Don’t roll.

If you choose IN the first round and then I will choose Don’t Roll at first. I will get \$14 but then after that I will choose roll each time after the 1st role. Chances are most likely you will get \$12 and I will get only \$10.I will the only take 7 rolls for you to get even with me. That way we both leave with a good amount of money. Hope you have a great evening and that this works out for both of us.	E	P	16(67%)
--	---	---	---------

– If you choose IN, and I roll, the chances of our getting the most \$ are very high. The likelihood of my rolling a 1 is small compared to the chances of rolling a 2–6. So we both get cash.	E	P	13(54%)
--	---	---	---------

Note: E: classified as “Empty Talk”; P: classified as “Promise”

is somewhat smaller but still significant. In particular, “weak promises” increase efficiency from 33% to 59% ($Z = 1.61$, $p = 0.05$, one-tailed test) and “strong promises” increase efficiency less, from 38% to 58% ($Z = 1.27$, $p = 0.10$, one-tailed test).

The results from the XH classifications contrast with both the results reported by C&D and the content analysis. With XH, we find that “weak promises” have a positive effect on efficiency, increasing it from 36% to 55%, but that this increase is statistically insignificant ($Z = 1.1$, $p > 0.10$, one-tailed test). On the other hand, “strong promises” are found to increase efficiency from 39% to 58%, a statistically significant increase ($Z = 1.3$, $p = 0.10$, one-tailed test).⁹

3.5 Discussion

We compared classifications of C&D’s (2006) messages among three procedures: C&D’s self-classification, content analysis, and the XH coordination game. Using content analysis, the classifications and implications of the classifications lined up reasonably well with those reported by C&D. However, little difference emerged between classifications in the “Weak” and “Strong” treatments, and those differences that did appear were counter-intuitive (more “strong” than “weak” promises, while the former category is in principle a subset of the latter).

We found significant differences in classifications between content analysis and the XH method. These differences are consistent with our hypothesis that coders are more responsive to classification criteria when their decisions are saliently rewarded. In our case, classifications with the XH game allow us to differentiate the impact of weak or strong promises, while classifications using content analysis do not.

It is worth emphasizing that differences we discovered between the XH game and content analysis may be sensitive to how we “trained” the coders to evaluate the messages. In particular, in light of our primary goals for this paper, and in order to maintain maximal symmetry with the coordination game approach (where there is no training), we adopted “minimal” training for our content analysis. More specifically, we simply gave our coders written instructions regarding what a promise meant (in two versions, both more detailed than anything received in the coordination game). Nevertheless, there clearly might be a way to train coders so that their responses line-up more closely with the results of our coordination game.

Differences between C&D and the XH coordination game classifications were also apparent: our evidence is that evaluators in our “weak” treatment used different criteria for “promise or intent” than C&D had in mind. This occurred despite the fact that those coders were given the classification criteria that appears in the C&D article. This highlights a limitation of any self-classification procedure: it can be very difficult to convey to others exactly how one generated the classifications. A significant advantage of non-subjective approaches, therefore, is that their results are more easily defended.

Using XH, the “Strong Promise” treatment leads to classifications that line up well with those provided by C&D, thus providing evidence that strong promises can

⁹We report one-tailed tests at the 10% significance level in order to be consistent with C&D, who reported one-tailed tests at the 10% significance level in their original contribution.

promote economic efficiency. This is consistent with a second paper by Charness and Dufwenberg (2010) that addresses whether watered-down promises are able to promote economic efficiency. Their conclusion is that they cannot, so “full-blooded” promises are necessary to foster cooperative economic exchange. It is worth emphasizing that we were able to reach the C&D (2007) result using C&D (2006) data, while neither content analysis nor self-classifications are able to do so.

4 Conclusion

We here described and implemented a laboratory game, originally introduced by Xiao and Houser (2005), to generate classifications of natural language messages. Scholars have developed a variety of algorithms for classifications of behavioral decision strategies (see, e.g., Houser et al. 2004, and cites therein), but these approaches do not easily extend to the domain of natural language. The game we described in this paper might be useful to anyone interested in exploring the role of language and communication in social and economic exchange, especially when investigators are either unwilling or unable to adopt explicit rules linking the words in a message to that message’s meaning.

To shed light on mechanisms underlying communication effects, ideally we would like to collect data on messages’ “true” intended meaning as well as “true” interpretations by targets of messages. Such data, unfortunately, are difficult (perhaps impossible) to acquire. Self-reports are not useful for this purpose. A reason is that it is unclear how to provide incentives for truthful revelation of either meaning or interpretation (e.g., subjects report meanings or interpretations to justify their actions, independent of their actual views).

The classification coordination game described here offers a procedure to test hypotheses regarding both the impact of types of messages on target subjects, as well as how a message writer’s decisions change according to the way she expects her target to interpret her message. A coordination game is useful for these purposes because in typical experiments on communication subjects write messages to strangers. The average opinion of a large number of evaluators, also strangers to the message writer, is a reasonable way to infer both how the message was likely interpreted, as well as the way in which the message writer expected the message to be interpreted. The XH classification game generates this information.

The XH game is a saliently-rewarded procedure for message classification, and we have demonstrated that this has advantages over traditional content analysis. There are also potential disadvantages. For example, our implementation of the XH classification game involved 25 coders, while traditional content analysis uses as few as two. Of course, XH does not necessarily require a large number of coders: how well the method would perform with fewer numbers is an open empirical question.

In order to provide incentives for classifiers, XH may require greater cost than traditional content analysis.¹⁰ Of course, the payment schedule in our design is just

¹⁰Note that in our experiment, in order to maintain symmetry between content analysis and the XH game, the fixed rate we paid coders is designed to be close to the average earnings of the coders in XH game.

one way to provide incentivized evaluations; many alternatives are available. For example, one might recruit only two coders (as in traditional content analysis) and randomly select two messages to determine their payment.¹¹ It is of course an open question whether and how classification outcomes may depend on the nature of the incentives.

Another potential drawback is that when there are a large number of messages XH may become impractical (e.g., thousands of messages with dozens of categories.) One complication in such environments is that it may not be feasible to recruit coders to come to the lab and perform classifications concurrently. It is worth noting, however, that the XH classification game does not necessarily require simultaneous participation. Coders may work on classifications outside the lab and at several different time periods, so long as anonymity among coders is maintained (ensuring that coordination is on message content). A potential drawback of the sequential approach, of course, is additional time.

Very little is known about “optimal” ways to classify natural language messages and further research in this area would be especially valuable. The XH classification game is easily replicable, and we have pointed to evidence that it works well. In particular, our results show that message classification with the XH game can inform hypotheses that cannot be addressed using either self-classifications or content analysis.

Acknowledgements We thank the editor, Tim Cason, and two anonymous referees for especially helpful comments on this manuscript. A grant to the first author from the International Foundation for Research in Experimental Economics funded this research. We thank Colin Camerer, John Dickhaut, Daniel Schunk, Vernon Smith and Jingjing Zhang for valuable feedback. We appreciate Jason Aimone’s able assistance with writing instructions, running experiments and organizing the data. We thank Kail Padgitt and Dorina Tila for assistance with experiments. We are especially grateful to Gary Charness and Martin Dufwenberg for their constructive comments on several earlier drafts of this paper. Any remaining shortcomings are our responsibility.

References

- Ben-Ner, A., Putterman, L., & Ren, T. (2007). *Lavish returns on cheap talk: non-binding communication in a trust experiment*. Mimeo.
- Brandts, J., & Cooper, D. (2007). It’s what you say, not what you pay: an experimental study of manager-employee relationships in overcoming coordination failure. *Journal of the European Economic Association*, 5(6), 1223–1268.
- Camerer, C. (2003). *Behavioral game theory*. Princeton: Princeton University Press.
- Cason, T., & Mui, V.-L. (2007). Communication and coordination in the laboratory collective resistance game. *Experimental Economics*, 10, 251–267.
- Cason, T., & Mui, V.-L. (2010). *Coordinating resistance through communication and repeated interaction*. Working paper, Purdue University.
- Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6), 1579–1601.
- Charness, G., & Dufwenberg, M. (2010). Bare promises. *Economics Letters*, 107, 281–283.
- Cooper, D., & Kagel, J. (2005). Are two heads better than one? Team versus individual play in signaling games. *American Economic Review*, 95(3), 477–509.
- Corazzini, L., Kubez, S., Maréchal, M.A., & Nicolóx, A. (2009). *Elections and deceptions: theory and experimental evidence*. Working paper.

¹¹We thank an anonymous referee for this suggestion.

- Crawford, V. (1997). Theory and experiment in the analysis of strategic interactions. In D. Kreps & K. Wallis (Eds.), *Advances in economics and econometrics: theory and applications. Seventh world congress* (Vol. I). Cambridge: Cambridge University Press.
- Dawes, R. M., McTavish, J., & Shaklee, H. (1977). Behavior communication and assumptions about other people's behavior in a commons dilemma situation. *Personality and Social Psychology*, 35(1), 1–11.
- Ellingsen, T., & Johannesson, M. (2007). Anticipated verbal feedback induces altruistic behavior. *Evolution and Human Behavior*. doi:10.1016/j.evolhumbehav.2007.11.001.
- Engle-Warnick, J., & Soroka, S.N. (2009). Harnessing the power of focal points to measure social agreement. Working paper.
- Epstein, J. (2007). *Generative social science*. Princeton: Princeton University Press.
- Feiler, L., & Camerer, C. (2009). Code creation in endogenous merger experiments. *Economic Inquiry*, 48(2), 337–352.
- Fouraker, L., & Siegel, S. (1963). *Bargaining behavior*. New York: McGraw-Hill.
- Hennig-Schmidt, H., Li, Z.-Y., & Yang, C. (2008). Why people reject advantageous offers—non-monotone strategies in ultimatum bargaining. *Journal of Economic Behavior and Organization*, 65, 373–384.
- Holt, C. (1995). Industrial organization: a survey of laboratory research. In J. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics*. Princeton: Princeton University Press.
- Houser, D., & Wooders, J. (2006). Reputation in auctions: theory, and evidence from eBay. *Journal of Economics and Management Science*. doi:10.1111/j.1530-9134.2006.00103.x.
- Houser, D., Keane, M., & McCabe, K. (2004). Behavior in a dynamic decision problem: an analysis of experimental evidence using a Bayesian type classification algorithm. *Econometrica*, 72(3), 781–822.
- Issac, R. M., & Walker, J. (1988). Communication and free riding behavior: the voluntary contributions mechanism. *Economic Inquiry* 26. doi:10.1111/j.1465-7295.1988.tb01519.x.
- Kimbrough, E., Smith, V., & Wilson, B. (2008). Historical property rights, sociality, and the emergence of impersonal exchange in long distance trade. *American Economic Review*. doi:10.1257/aer.98.3.1009.
- Krippendorff, K. (2004). *Content analysis: an introduction to its methodology* (2nd edn.). Thousand Oaks: Sage.
- Luce, R. D., & Raiffa, H. (1957). *Games and decisions: introduction and critical survey*. New York: Wiley.
- Ledyard, J. O. (1995). Public goods: a survey of experimental research. In J. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics*. Princeton: Princeton University Press.
- Ochs, J. (1995). Coordination problems. In J. Kagel & A. E. Roth (Eds.), *The handbook of experimental economics*. Princeton: Princeton University Press.
- Schotter, A., & Sopher, B. (2007). Advice and behavior in intergenerational ultimatum games: an experimental approach. *Games and Economic Behavior*, 58, 365–393.
- Smith, V. (1962). An experimental study of competitive market behavior. *Journal of Political Economy*, 70(2), 111–137.
- Sutter, M., & Strassmair, C. (2009). Communication, cooperation and collusion in team tournaments—an experimental study. *Games and Economic Behavior*, 66, 506–525.
- von Ahn, L. (2005). Games with purpose. PhD Dissertation, Carnegie Mellon University.
- von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton: Princeton University Press.
- Xiao, E., & Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, 102(20), 7398–7401.
- Xiao, E., & Houser, D. (2007). *Emotion expression and fairness in economic exchange*. Manuscript, George Mason University.