Mediated Talk: An Experiment[*]

Andreas Blume[†] Ernest K. Lai[‡] Wooyoung Lim[§]

November, 2022

Abstract

We experimentally compare mediated (cheap) talk with direct (cheap) talk. Theory, guided by a characterization of equilibria in both environments, suggests that mediated talk has the potential to improve information sharing and welfare relative to direct talk. We sharpen the theory prediction by invoking Crawford's [24] language-anchored level-k analysis. In the experiment, we find that mediated talk can indeed facilitate information transmission. We also find, however, that this requires that the language employed conforms with the mediation mechanism: mediation mechanisms improve information sharing for a variety of conforming languages, but fail to do so with a nonconforming language. These experimental findings match the predictions from the language-anchored level-k analysis. Strikingly, this is the case even when a whole array of alternative selection criteria (including iterative deletion of dominated strategies, strict equilibrium, Pareto efficiency, etc.) make a unique common prediction that sharply disagrees with the language-anchored level-k prediction.

Keywords: Sender-Receiver Games, Communication, Mediation, Noisy Channels, Language, Laboratory Experiments

***JEL* classification numbers**: C72; C92; D82; D83

[*] We are grateful to Alessandra Casella, Vince Crawford, Shih En Lu, Guillaume Fréchette, Marta Serra-Garcia, Joel Sobel, and Tiemen Woutersen for helpful suggestions. We have benefitted from comments by seminar audiences at the Asian Meeting of the Econometric Society at Xiamen University, Claremont Graduate University, the Conference on Strategic Information Transmission at Peking University, the European Meeting of the Economic Science Association (ESA), Lingnan University, New York University, the Society for the Advancement of Economic Theory (SAET) International Conference in Ischia, Shanghai University of Finance and Economics, Simon Fraser University, the Southwest Economic Theory (SWET) conference, the University of Calgary, the University of Central Florida, and the Virtual East Asia Experimental and Behavioral Economics Seminar Series. This study is supported by a grant from the Hong Kong Research Grant Council (Grant ECS-699613).

[†] Department of Economics, University of Arizona. *Email:* ablume@arizona.edu

[‡] Department of Economics, Lehigh University. *Email:* kwl409@lehigh.edu

[§] Department of Economics, The Hong Kong University of Science and Technology. *Email:* wooyoung@ust.hk

1 Introduction

Information transmission is a ubiquitous part of economic activity and inefficiencies affecting it entail potentially significant costs to society. A principal source of inefficiencies is the incentive for strategic manipulation resulting from divergence of interests among communicating parties (Crawford and Sobel [23]). The extent of these inefficiencies depends on the rules and protocols that govern communication. In this paper we engage in a *communication design* exercise: we investigate whether introducing a (non-strategic) mediator can improve information transmission.

Various authors propose ways for organizations to improve internal information transmission by implementing communication protocols that mitigate misreporting. Harris and Raviv [34] suggest the use of intermediaries, like the board of directors, to achieve better communication between management and shareholders. Laclau, Renou and Venel [42] demonstrate the benefits from requiring reporting through multiple channels and use it to rationalize the matrix organization as a management structure. Ambrus, Azevedo, Kamada, and Takagi [2] look at legislative committees as information intermediaries and show, building on work by Ivanov [37], that with strong misalignment of interests a biased committee can improve information transmission between privately informed lobbyists and the legislature.

A common feature of the proposed protocols is that they enable the garbling of information. With a non-strategic intermediary that garbling can be achieved directly. With a strategic intermediary, it may be the result of (and require) the intermediary using a mixed strategy (see Ambrus, Azevedo, and Kamada [3]). There is little experimental work that would shed light on whether these garbling schemes are effective and on how to make them more effective. We see this paper as a beginning of a systematic investigation of communication protocols that mitigate misreporting incentives by facilitating the garbling of information. We focus on the most direct implementation of garbling in sender-receiver games, through a non-strategic mediator.

The rationale for the benefits of garbling in sender-receiver games is straightforward: A sender who is reluctant to provide information will be more willing to do so if it is known that information is degraded by garbling. The receiver prefers garbled information to no information. Under the right conditions both parties gain.[1]

The potential for mediation to improve information transmission has long been recognized. Forges [29] presents an example of an information transmission game in which communication is entirely ineffective with direct communication, while both sender and receiver can gain from communication via an appropriately chosen mediation scheme. The underlying logic is nicely illustrated by Myerson [47] with a story of a sender and receiver communicating via a messenger pigeon. The

[1] The sender could accomplish the garbling of messages without a mediator by adopting an appropriate randomization. The problem is that without a mediator that randomization will typically not be incentive compatible. The mediator has a role because he is committed to a garbling rule.

sender has the choice of either sending the pigeon or not sending the pigeon. If the pigeon is sent, it gets lost with some probability. There are two sender types, one who prefers to be revealed and another who prefers to be concealed. There is an equilibrium in which the type who prefers to be revealed sends the pigeon and the other does not send the pigeon. When the pigeon does not arrive, the receiver does not know whether it was never sent or was sent but got lost. As a result, when the pigeon does not arrive the receiver remains uncertain about whether he is dealing with one type or the other. The type who prefers to be concealed remains concealed, and thus achieves deniability, whereas the type who prefers to be identified manages to get identified at least some of the time.

More recently Goltsman, Hörner, Pavlov and Squintani [31] have taken up the question of mediation in the context of the leading example in Crawford and Sobel [23]. They identify an efficiency bound for optimal mediation. Via the revelation principle (Myerson [46]) this is also the bound for any other communication protocol, including, for example, repeated face-to-face communication as considered by Krishna and Morgan [41]. Blume, Board and Kawamura [9] demonstrate that the efficiency bound can be attainted in a single round of communication through a simple noisy channel: the sender sends a message to the receiver that goes through with some probability as sent; otherwise the message is replaced by a random draw from some fixed distribution. The equilibria that achieve the efficiency bound with this noisy channel exhibit a structure reminiscent of the messenger pigeon example: (sets of) high types are sometimes pooled with (a set of) low types, and otherwise revealed.

We take mediation to the lab. To our knowledge, this is the first paper that experimentally compares direct with mediated cheap-talk communication. To keep the comparison manageable and induce salient incentives, we use a two-type incentive structure and a mediation rule that closely mirrors the one employed in the messenger pigeon example.

Our objective is twofold. First, we are interested in comparing the outcomes from direct talk, where the receiver observes the sender's message as sent, and mediated talk, where the sender's message is filtered through a noisy channel. Second, and intimately related, we are interested in how the language that is available to subjects affects mediation outcomes, where by "language" we refer to the framing of messages in our experiment. The question of whether mediation can improve on direct communication is a mechanism design problem. The language that is employed is part of the mechanism. It affects whether the mechanism is direct or indirect, whether desired outcomes can be supported with equilibria that require truth-telling, and whether desired outcomes can be supported with equilibria that require players to be obedient.

In an equilibrium analysis of communication games, the choice of language makes no difference at all (assuming that there is a one-to-one mapping between languages). To obtain sharper predictions, we complement and refine the equilibrium perspective with a level-k analysis. In the level-k analysis the language is central because, through pinning down level-0 behavior, it provides the anchoring of the level-k hierarchy. Given the multiplicity of equilibria in our games, both with and without mediation, it is interesting to know whether language affects equilibrium selection.

Most of the prior experimental literature on direct talk frames messages in terms of payoff types or sets thereof. Framing messages in terms of actions is a natural, under-explored, alternative. With mediation, there are additional concerns: (i) the language used for messages sent to the mediator may or may not match the language used for messages received from the mediator; (ii) the language determines whether the mechanism is direct or indirect; and (iii) languages admitting truthful behavior may either *conform with the mechanism*, by admitting truthful equilibria, or may not have truthful equilibria, and thus fail to conform with the mechanism.

For the sake of experimental control, and to prevent ballooning sizes of strategy spaces, the languages of our experiment are primitive. They lack almost all distinguishing features of human language: there is no syntax; the semantics is not productive; there is no generative structure (see, for example, Chierchia and McConnell-Ginet [22]). All our languages provide are minimal semantics that make it plausible for messages to refer to types, in one case, and to actions, in the other.

A natural frame to adopt with a mediator is that of a direct mechanism. In a direct mechanism the sender makes reports about her type to the mediator, using a language consisting of *declaratives* (declarations of her types). Based on those reports, the mediator makes action recommendations to the receiver, using a language consisting of *directives* (directions of which action to take).[2] In contrast, with direct communication the spaces of sent and received messages are identical.

Accordingly, we consider five different mediated-talk scenarios: (i) a *Mediated Direct Mechanism*, as described above; (ii) a mechanism that uses directives for both inputs and outputs, which we call *Mediated Directives*; (iii) a mechanism that uses declaratives exclusively, and has the language conform with the mechanism, which we term *Mediated Declaratives*; (iv) a mechanism that also exclusively uses declaratives, but so that the language does not conform with the mechanism, which we refer to as *Nonconforming Mediated Declaratives*; and, (v) a mechanism in which the declaratives language conforms with the mechanism but the garbling rate is insufficient to render mediation effective, which we refer to as *Mediated Declaratives with Excess Accuracy.* We compare these with each other and with the two direct talk scenarios, *Direct Directives*, in which messages are framed as directives, and *Direct Declaratives*, in which messages are framed as declaratives.

We obtain our theoretical predictions from augmenting a standard equilibrium analysis with a

[2] Jakobson [38] identifies six functions of language: emotive (aimed at expressing the speaker's emotion), poetic (focused on the message for its own sake), phatic (aimed at establishing, prolonging, or discontinuing communication), metalingual (aimed at verifying the use of a common code), referential (oriented toward the referent), and conative (oriented toward the addressee). Lewis [43] notes that meaning can be conceived of as "a signal that a state of affairs holds" or, alternatively, as "a signal to do something"; a signal can be "indicative" or "imperative." What we call "declaratives" and "directives" mirrors the referential and conative functions of language according to Jakobson and Lewis's distinction of signals-that from signals-to. Bühler [13], who inspired Jakobson, distinguishes "Ausdruck" (expression), "Darstellung" (representation), and "Appell" (appeal). The latter two correspond to Jakobson's referential and conative functions of language, and resonate with our declaratives and directives.

level-k approach that, following Crawford [24], is anchored at the focal meanings of messages: level-0 senders are taken to be forthright (i.e., truthful with a declaratives language and indicating their preferred response with a directives language) and level-0 receivers to be credulous (in the sense of best-responding to level-0 senders, following Crawford [24]); higher-level behavior is obtained by iterating best replies. The level-k prediction refines the equilibrium prediction. For direct talk, the equilibrium analysis by itself, without invoking level-k reasoning, predicts pooling, which can be achieved with having messages be independent of the sender's type. With mediated talk, the equilibrium analysis by itself allows for both separation, where the two types send distinct messages, and pooling. The level-k analysis is consistent with the equilibrium analysis: at all levels above level 0 players use equilibrium strategies. The level-k prediction refines the equilibrium prediction: For direct talk it singles out exactly one message that will be sent. For mediated talk it predicts separation with conforming languages and pooling on a particular message with a non-conforming language.

We find that, in line with these theoretical predictions, the modal observed behavior of both senders and receivers converges to pooling with direct talk, to separation with mediated talk with conforming languages, and to pooling with mediated talk when the language is non-conforming. The agreement with the theoretical predictions extends to the details of behavior, notably message use. The difference in behavior between direct talk and mediated talk with a conforming language translates into payoff advantage of the latter over the former.

There is a stark difference in observed behavior under mediation depending on whether the language does or does not conform with the mechanism. Mediation mechanisms improve information sharing for a variety of conforming languages, but fail to do so with a nonconforming language. These experimental findings match the predictions from the language-anchored level-k analysis. Strikingly, this is the case even when a whole array of alternative selection criteria (including iterative deletion of dominated strategies, strict equilibrium, Pareto efficiency etc.) make a unique common prediction that sharply disagrees with the language-anchored level-k prediction.

There is a rich literature on sender-receiver game experiments with direct talk. We survey that body of work in Blume, Lai, and Lim [11]. The earliest paper in this stream is Dickhaut, McCabe, and Mukherji [26] who implement a discretized version of the Crawford-Sobel model. Much of this literature frames sender messages as reports of types or sets of types. An exception is Blume, DeJong, Kim, and Sprinkle [10], who consider messages that are framed as action recommendations. In the present paper, one of the questions we ask is whether the framing of messages, the choice of language, has an impact on the performance of mediation mechanisms.

The experimental papers most closely related to ours are Nguyen [48], Blume, Lai, and Lim [12], and Fréchette, Lizzeri, and Perego [30]. Nguyen and Fréchette et al. experimentally investigate

Bayesian persuasion (Kamenica and Gentzkow [40]).[3] With Bayesian persuasion, like in the present paper, the receiver observes a garbled signal of the state of the world; unlike here, there is no private information and the sender can fully commit to a signaling rule that maps states of the world into signals.[4] Blume et al. [12] model and implement Warner's [52] randomized response method in the lab. Under randomized response, garbling is entirely under the control of the sender, and for that to be incentive compatible it is necessary that the sender has a preference for compliance with the procedure, which may be in the form of deriving utility from truth-telling. While messages in the present paper are cheap talk, communication under randomized response amounts to costly signaling. Ours is the first paper that looks at the effects of garbling cheap-talk messages in sender-receiver games.

Casella, Friedman, and Perez [16] experimentally study mediation in a two-player conflict resolution game based on theoretical work by Hörner, Morelli, and Squintani [35]. Both players have private information, send messages, and take actions after the exchange of messages. Like in the present paper garbling of messages offers the promise of efficiency gains, in this case through a reduction in conflict. They find that mediation significantly affects behavior at the communication stage but does not reduce conflict.

Chassang and Zehnder [18], building on Chassang and Padró i Miquel [17], use an experiment to investigate the role garbling in encouraging whistleblowers to come forward. In their environment exogenous garbling of messages is predicted to help create a deterrent: if whistleblowers' reports of observed misbehavior are received even when not intended, committing to retaliate becomes costly for the misbehaving party, whereas without garbling commitment to retaliation can be an effective off-path threat. In their experiment, Chassang and Zehnder [18] find support for this prediction.

Our paper intersects with the broader literature on truth-telling and obedience in mechanism design. For some of the mechanisms we consider, messages to the mediator can be viewed as type reports. For others, messages from the mediator can be viewed as action recommendations. In our direct mechanism, we have both type reports and action recommendations. When messages are type reports, we can look at truth-telling behavior of senders. When they are action recommendations, we can ask whether receivers are obedient. The experimental literatures on using VCG mechanisms to deal with public goods problems (surveyed in Chen and Ledyard [20]) and on using strategy-proof mechanisms to address school choice (e.g. Chen and Sönmez [21], and the survey by Hakimov and Kübler [33]) have examined truth-telling behavior when it is a dominant strategy. The experimental literature on implementing correlated equilibria (e.g., Duffy and Feltovitch [27]) has asked whether players are obedient when given action recommendations that are exogenously generated from

[3] Fréchette et al. [30] allow for different degrees of commitment and thus are positioned on the continuum between cheap talk and Bayesian persuasion.

[4] Au, Kwon and Li [4] incorporate reciprocity into the Bayesian persuasion environment and show that reciprocity concerns lead the sender's optimal persuasion strategy to be more informative. Their main theoretical findings are confirmed by the experimental data.

a correlated equilibrium distribution. We add to these perspectives, by (i) having truth-telling incentives (of senders) be contingent not only on the mechanism, but also on the behavior of receivers, and (ii) by having obedience incentives (of receivers) not only dependent on the mechanism and other players' actions, but also on the information revealed by senders. Neither senders nor receivers in our mechanisms have dominant strategies.

Our interest in language in mechanism design has antecedents in both the literature on auctions and the one on school choice. Masatlioglu, Taylor, and Uler [44] experimentally compare direct with indirect mechanisms by varying the language through which bidders generate their bids in a first-price auction. Bichler, Milgrom and Schwarz [7] is a recent contribution to the literature on bidding languages in combinatorial auctions. They propose and evaluate an alternative to the widely used enumerative exclusive or (XOR) bidding language. Bichler, Goeree, Mayer and Shabalin [8] use a laboratory experiment to compare simple bid languages with more expressive bidding languages. Calsamiglia, Haeringer and Klijn [15] compare school choice mechanisms in which parents have access to a language that lets them fully express their ranking of schools with alternative languages in which parents are constrained to list a limited number of schools.

In the next section we introduce the communication protocols and the incentive structure. In Section 3 we discuss the theoretical predictions. Section 4 describes our experimental treatments and procedures. In Section 5 we report our findings and in Section 6 we discuss our findings and possible extensions.

2 The Communication Environments

We investigate the impact of mediation on communication between a privately informed sender and an uninformed receiver whose action determines the payoff of both players. In mediated talk the sender sends a message to a non-strategic mediator who in turn sends a message to the receiver. We contrast mediated talk with direct talk, where the sender sends a message to the receiver without the intervention of a mediator. In both cases, messages have no direct effect on payoffs, and are thus cheap talk.

	L	C	R
s	$110, 120$	$10, 0$	$60, 100$
t	$80, 0$	$10, 120$	$130, 90$

Table 1: Payoffs

The payoff and information structures are the same across all environments we consider. There are two possible states of the world, s and t, which are commonly known to be equally likely. The receiver has three actions L, C and R. Table 1 displays payoffs as a function of the state and the receiver's action. In each cell of the payoff table the first entry indicates the sender's payoff and the second entry the receiver's payoff. This incentive structure captures two central features of the one made prominent by Crawford and Sobel [23]: there are possible efficiency gains from communication and there are incentives for the sender to misrepresent her type. Without communication, the receiver's unique optimal action is R. Both sender and receiver could gain if the sender could credibly reveal when her type is s. There is, however, an incentive problem because if type s credibly reveals, then type t prefers to mimic s rather than be revealed herself.

Table 2: Communication Environments

Game	Direct talk	Mediated talk
Transmission rule (input $\overset{\text{prob.}}{\longrightarrow}$ output)	$m_1 \xrightarrow{1} m_1$ $m_2 \xrightarrow{1} m_2$	$\begin{tikzcd} m_1 \arrow[r, "p=\frac{1}{2}"] \arrow[dr, "1-p=\frac{1}{2}"] & \widetilde{m}_1 \\ m_2 \arrow[r, "1"] & \widetilde{m}_2 \end{tikzcd}$

Table 2 summarizes the transmission rules in direct and mediated talk for generic messages, $m_1, m_2, \widetilde{m}_1$, and $\widetilde{m}_2$. With direct talk sent messages coincide with received messages. Mediated talk differs from direct talk in two ways: (i) The set of messages available to the sender $\{m_1, m_2\}$ may differ from the set of messages $\{\widetilde{m}_1, \widetilde{m}_2\}$ that may be transmitted by the mediator to the receiver – as it will be the case when the mediator is a direct mechanism (Myerson [46]). (ii) In addition, the transmission rule is stochastic; our primary focus is on a transmission rule in which message m_1 is mapped into a distinct message $\widetilde{m}_1$ with probability $p = 1/2$ and otherwise is mapped into the message $\widetilde{m}_2$, the same message into which message m_2 is mapped.

We refer to the probability p as the "accuracy" of the mechanism. An optimal mechanism (as well as the optimal Bayesian persuasion scheme) would have an accuracy of $p = 7/10$ instead of 1/2. In an optimal mechanism the receiver would be indifferent between actions C and R following message $\widetilde{m}_2$ from the mediator. We choose to avoid this indifference in order to help making incentives salient in the experiment. Mechanisms with a higher accuracy than $p = 7/10$ are not incentive compatible. With that in mind, we do consider $p = 9/10$ as a control.

With both mediated and direct talk, we need to choose a language, i.e., how we label the messages $m_1, m_2, \widetilde{m}_1$, and $\widetilde{m}_2$. To give effective communication its best chance and because we want to allow for direct mechanisms, we use languages that relate to either types or actions, rather than generic languages. With direct talk, we are constrained to use languages in which sender and receiver labels are identical; with mediated talk they may differ. With mediated talk, in addition to the set of labeled messages available to the sender, we also have to be concerned with which of the labels attaches to m_1 and which to m_2. As we will see, for a fixed set of labeled messages available to the sender, it is possible that for one attachment there is an equilibrium in which the sender is forthright, whereas switching the attachments rules out forthright equilibria. We say that in the former case the *language conforms with the mechanism* and in the latter it does not.[5]

Our language choices are guided by the following desiderata: (i) For mediated talk, we want to compare direct with indirect mechanisms. (ii) We want to compare direct with mediated talk while keeping the language constant. (iii) With direct talk and indirect mechanisms, we want to use languages that are necessitated by direct mechanisms. (iv) With mediated talk, we want to compare the performance of conforming and non-conforming langauges. As we will see, nonconforming languages are of interest because, while they preserve joint distributions over types and actions that can be achieved in equilibrium, they reveal a striking contrast between level-k predictions and predictions obtained from a host of alternative equilibrium selection rules.

To satisfy (i), we consider a direct mechanism in which m_1 = "s", m_2 = "t", $\widetilde{m}_1$ = "L", and $\widetilde{m}_2$ = "R", i.e., senders send reports of their types to the mechanism and receivers obtain action recommendations from the mechanism (we use quotes in the paper to distinguish type reports from types and action recommendations from actions). We refer to the corresponding treatment as *Mediated Direct Mechanism.* Indirect mechanisms that make minimal changes to this language use either the language in which $m_1 = \widetilde{m}_1$ = "s" and $m_2 = \widetilde{m}_2$ = "t", which we refer to as *Mediated Declaratives*, or the language in which $m_1 = \widetilde{m}_1$ = "L" and $m_2 = \widetilde{m}_2$ = "R", which we refer to as *Mediated Directives.* As a control we also examine a version of Mediated Declaratives in which the transmission accuracy p is raised from $p = 1/2$ to $p = 9/10$; we refer to this treatment as *Mediated Declaratives with Excess Accuracy.* Given these indirect mechanisms, we can satisfy desideratum (ii) with either the *Direct Declaratives* language, in which m_1 = "s" and m_2 = "t", or the *Direct Directives* language, in which m_1 = "L" and m_2 = "R". To satisfy (iii) we consider both declaratives and directives languages. Finally, in order to satisfy (iv), we examine the language in which $m_1 = \widetilde{m}_1$ = "t" and $m_2 = \widetilde{m}_2$ = "s", which we refer to as *Nonconforming Mediated Declaratives.*

3 Theoretical Predictions

For our predictions, we proceed in two steps. We first fully characterize the set of equilibrium outcomes (defined as joint distributions over sender types and receiver actions) for each game. Second, since this characterization does not address multiplicity of equilibria and is silent about the specifics of message use, we supplement it by a level-k analysis anchored in language. Anchoring predictions in the available language gives us a unique prediction for each game.

We follow Crawford's [24] proposal for how to handle communication games with focal message meanings in a level-k framework (see also Cai and Wang's [14] application to Crawford-Sobel type

[5] Similar considerations apply to receiver labels. We do not pursue those in this paper.

sender-receiver games and Wang, Spezio, and Camerer's [51] support for the level-k model in sender-receiver games via eyetracking). The key is that in communication games with focal message meanings those meanings are a natural anchor for players' reasoning. Specifically, we model level-0 (L_0) senders as being *forthright* and L_0 receivers as being *credulous.* A forthright sender is truthful when messages are framed as declaratives and recommends her preferred action when messages are framed as directives. A credulous receiver best-responds to a forthright sender.[6] For higher levels of sophistication, we have $L_{k\geq 1}$ senders (levels k that are at least $k = 1$) best-respond to L_{k-1} receivers and $L_{k\geq 1}$ receivers best-respond to L_k senders. We further postulate that (i) receivers who encounter an unexpected message behave as they would at the next lower level, and (ii) senders who have multiple best replies retain their strategy from the next lower level, provided that it is one of those best replies (these choices are somewhat arbitrary and worth revisiting upon examining the data). We refer to the predicted behavior of $L_{k\geq 1}$-players (thus excluding level-0 behavior) as the prediction from the level-k analysis.[7]

As we will see, in each game the set of equilibrium outcomes is small. The level-k analysis refines this prediction further and selects a single equilibrium for each case.

3.1 Direct Talk

With direct talk, *separation*, where different types of the sender send distinct messages, is not part of an equilibrium. Type t of the sender would receive her lowest possible payoff, 10, with separation. She would be better off mimicking type s for a payoff of 80. This breaks any candidate for a separating equilibrium.

	L	C	R
s	0%	0%	**50%**
t	0%	0%	**50%**

Table 3: Pooling

Like in any sender-receiver game, with direct talk *pooling* is supported by an equilibrium. Under

[6] In calling a receiver who best-responds to a level-0 sender "credulous," we adopt the terminology of Crawford [24].

[7] Our setup introduces two aspects that are potentially relevant for a level-k analysis and not present in Crawford [24]: Messages are sometimes framed as directives or garbled by a mediator. As a result, with directives even an unsophisticated sender needs to pay attention to payoffs if we want to connect her message use to focal message meanings. In addition, with directives rather than expressing receiver credulity by having L_0 receivers best-respond to a forthright sender strategy, we could have L_0 receivers take directives at face value. Finally, a decision has to be made on how credulous receivers, if they best-respond to forthright senders, deal with mediation. Our modeling choices are motivated by trying to stay close to Crawford's original formulation and maximizing the predictive power of our level-k analysis.

pooling, regardless of the state of the world, the receiver takes the action that is optimal given his prior beliefs, here action R. The pooling equilibrium outcome is shown in Table 3. It reflects the fact that types are equally likely and that the receiver takes action R irrespective of the type. We show in the appendix that ***under direct talk pooling is the unique equilibrium outcome.***

The equilibrium analysis remains silent about how messages will be used in equilibrium under direct talk. There are equilibria in which the sender sends messages s regardless of type, equilibria in which she sends messages t regardless of type, as well as a continuum of equilibria with different forms of randomization. The level-k analysis can be more precise because it makes use of focal message meanings.

Table 4: Level-k Prediction for Direct Declaratives

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"t"	L	C
$L_{k\geq 1}$	"s"	"s"	R	C

Table 4 summarizes the level-k prediction for Direct Declaratives. With the message space $\{$"s", "t"$\}$, forthright L_0 senders are truthful. Credulous L_0 receivers take senders to be truthful and respond accordingly. Senders at level L_1, who best-respond to L_0 receivers strictly prefer to send message "s", regardless of their true type. Therefore, since message "s" is uninformative, L_1 receivers respond with action R. This pattern of behavior is stable and iterates through all higher levels: at all but the lowest level of sophistication the prediction is that senders send message "s", receivers respond with the pooling action R to message "s", and with action C to message "t".[8] Hence, the level-k analysis selects a unique pooling equilibrium: at all but the lowest level players employ the same pooling-equilibrium strategy. ***For observables in Direct Declaratives the level-k prediction for non-L_0 types is that the sender always sends message "s" and that the receiver responds to message "s" with action R.***

The level-k prediction for Direct Directives is summarized in Table 5. With message space $\{$"L", "R"$\}$ (forthright) L_0 senders ask for their preferred action. L_0 receivers trust that senders will be forthright and respond accordingly. At all levels other than the lowest level senders pool on message "L" and receivers respond to message "L" with action R and to message "R" with action C. ***For observables in Direct directives, the level-k prediction for non-L_0 types is that the sender always sends message "L" and that the receiver responds to message "L"***

[8] This level-k prediction is also supported by combining a monotonicity requirement with iterated deletion of dominated strategies, as proposed by Gordon, Kartik, Lo, Olszewski and Sobel [32]: suppose we order the type space such that $s > t$, the action space such that $L > R > C$, the message space such that "s" > "t", and restrict players to monotonic strategies, that is, the message of type s has to be weakly higher than the message of type t and the action following message "s" has to be weakly higher than the action following message "t"; then the unique strategy pair that survives iterated deletion of weakly dominated strategies is for the sender to send message "s" regardless of type, and for the receiver to respond to "s" with R and to "t" with C.

with action R.

Table 5: Level-k Prediction for Direct Directives

	Sender's Strategy		Receiver's Strategy	
	s	t	"L"	"R"
L_0	"L"	"R"	L	C
$L_{k\geq 1}$	"L"	"L"	R	C

3.2 Mediated Talk with a Conforming Language

In this section, we use "mediated talk" without qualifications to refer to "mediated talk with a conforming language." These, as we will see, are the languages with maximal promise for improving on direct talk. We discuss mediated talk with a nonconforming language in the next subsection.

With mediated talk, *pooling* continues to be supported by multiple equilibria. Unlike with direct talk, with mediated talk *separation*, the joint distribution over types and actions that arises when the sender uses a forthright separating strategy and the receiver best-responds to that strategy, is supported by an equilibrium. To see this, consider the mediated-direct-mechanism game (the argument applies to the other mediated-talk games with the appropriate relabeling of messages). When the sender reports truthfully, i.e., sends "s" to the mediator when her type is s and likewise sends "t" when her type is t, the receiver assigns posterior probability 1 to type s after receiving message "L" from the mediator and assigns posterior probability $\frac{1}{3}$ to type s after receiving message "R" from the mediator. This makes it a (unique) best reply for the receiver to take action L after message "L" from the mediator and to take action R after message "R" from the mediator. Given that strategy of the receiver, it is uniquely optimal for the sender to report truthfully.[9] The joint distribution over types and actions that arises from separation is shown in Table 6.

	L	C	R
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Table 6: Separation

We demonstrate in the appendix that ***with mediation, the only two equilibrium outcomes***

[9] The structure of this equilibrium is reminiscent of that of optimal equilibria in Blume, Board, and Kawamura [9]. They analyze communication through a noisy channel in the uniform-quadratic version of the Crawford-Sobel [23] model. Equilibria that attain the efficiency bound established by Goltsman, Hörner, Pavlov and Squintani [31] have intervals of high types be revealed some fraction of the time and otherwise pooled with the lowest interval of types. In both cases, this provides a cover for low types. Here the fact that s is sometimes pooled with t protects type t from receiving her least favorite action C.

are separation and pooling. Furthermore, ***there is a unique equilibrium supporting separation***, which implies that ***for the case of separation, the equilibrium analysis pins down message use.*** For the case of pooling, the equilibrium analysis does not pin down message use.

Separation under mediation is also the prediction obtained from a wide variety of equilibrium refinements: only the separating equilibrium survives iterative deletion of dominated strategies (independent of the order of deletion), and it is the unique Pareto efficient equilibrium, the unique strict equilibrium, the unique persistent equilibrium (Kalai and Samet [39]), and the unique equilibrium belonging to a minimal curb set (Basu and Weibull [5]). Thus, a host of ***equilibrium refinements predict*** **separation** ***with mediation.***

Table 7: Level-k Prediction for Mediated Declaratives

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"t"	L	R
$L_{k\geq 1}$	"s"	"t"	L	R

Table 8: Level-k Prediction for Mediated Directives

	Sender's Strategy		Receiver's Strategy	
	s	t	"L"	"R"
L_0	"L"	"R"	L	R
$L_{k\geq 1}$	"L"	"R"	L	R

Table 9: Level-k Prediction for Mediated Direct Mechanism

	Sender's Strategy		Receiver's Strategy	
	s	t	"L"	"R"
L_0	"s"	"t"	L	R
$L_{k\geq 1}$	"s"	"t"	L	R

Additional support for the separation prediction with mediation comes from the level-k analysis. Tables 7-9 report the level-k analysis for the three mediated-talk games. In all three cases forthrightness of the sender at the lowest level translates into separation, which is preserved through all levels.

The level-k analysis predicts that with mediated declaratives senders are truthful and receivers respond to message "s" with action L and to message "t" with action R, with mediated directives senders sincerely ask for their favorite action and receivers respond to message "L" with action L and to message "R" with action R, and with the mediated direct mechanism senders are truthful and receivers respond to message

“L” ***with action*** L ***and to message*** “R” ***with action*** R***.***[10]

3.3 Mediated Talk with a Nonconforming Language

Recall that *Nonconforming Mediated Declaratives* uses the language in which $m_1 = \widetilde{m}_1 =$ “t” and $m_2 = \widetilde{m}_2 =$ “s”. With that language, message “s” always gets transmitted as sent, whereas the probability that message “t” gets transmitted as sent is now 1/2.

As with a conforming language, Pareto efficiency, iterative deletion of dominated strategies, strictness, persistence, and the minimal curb criterion all select the same (separating) equilibrium. Unlike with a conforming language, in this equilibrium the sender inverts the forthright separating strategy: type s of the sender sends message “t” and type t of the sender sends message “s”. In stark contrast, the level-k analysis now predicts pooling.

Table 10: Level-k Prediction for Nonconforming Mediated Declaratives

	Sender's Strategy		Receiver's Strategy	
	s	t	“s”	“t”
L_0	“s”	“t”	R	C
$L_{k\geq 1}$	“s”	“s”	R	C

As shown in Table 10, ***with nonconforming mediated declaratives, the level-***k ***analysis predicts pooling, with both types of the sender exclusively using message*** **“*s*”*.*** The sharp divergence of the level-k prediction from the equilibrium refinement prediction is intriguing. It puts the level-k approach to the test when there is an alternative with strong credentials. Given the uniqueness of the refinement prediction and its attractive payoff properties, fully rational players might be able to intuit it and behave accordingly. Furthermore, since the separating equilibrium is the unique equilibrium that belongs to (in fact, constitutes) a minimal curb set, it is a natural target for adaptive learning (Hurkens [36]). Thus, even if players are boundedly rational, there might be forces that push them toward the separating equilibrium with repeated play.

3.4 Mediated Talk with Excess Accuracy

In mediated declaratives with excess accuracy, the language, which is given by $m_1 = \widetilde{m}_1 =$ “s” and $m_2 = \widetilde{m}_2 =$ “t”, is the same as for mediated declaratives, while the accuracy with which message “s” is transmitted is increased from $p = 1/2$ to $p = 9/10$.

[10] Our level-k analysis is sender-anchored, starting with level-0 senders who are forthright, and alternating best replies from that anchor. Alternatively, one could conduct a receiver-anchored level-k analysis, starting with level-0 receivers who believe declaratives and are obedient in response to directives. The predictions for levels 2 and higher coincide in all five games. Except for the direct-directives game, they also coincide for all levels 1 and higher.

With accuracy $p = 9/10$, the only equilibria are pooling equilibria – mediation is rendered ineffective. The level-k analysis is more specific and predicts that both sender types send message "s" exclusively. Hence the predicted sender behavior for mediated declaratives with excess accuracy is the same as for direct declaratives. For receivers, the predictions for mediated declaratives with excess accuracy and for direct declaratives differ. Senders in direct declaratives are predicted not to send message "t" and our level-k off-path postulate for receivers prescribes that receiver's respond with action C to the unsent message "t". In contrast, in mediated declaratives with excess accuracy even though senders are also predicted to send message "s" exclusively, because of the noisy channel, receiving message "t" is on path, and equally likely to have been triggered by a type s sender and a type t sender. Hence the receiver's unique best reply to observing message "t" is action R.

Table 11: Level-k Prediction Mediated Declaratives with Excess Accuracy

	Sender's Strategy		Receiver's Strategy	
	s	t	"s"	"t"
L_0	"s"	"t"	L	C
$L_{k\geq 1}$	"s"	"s"	R	R

As shown in Table 11, ***in mediated declaratives with excess accuracy, the level-k analysis predicts pooling, with both types of the sender exclusively using message "s" and the receiver responding to both messages with action R.***

4 Experimental Treatments and Procedures

Each of the seven games analyzed above corresponded to an experimental treatment, as summarized in Table 12.

Our experiment was conducted in English using z-Tree (Fischbacher [28]), in face-to-face mode, and oTree (Chen, Schonger, and Wickens [19]), in real-time online mode, at The Hong Kong University of Science and Technology. A total of 638 subjects participated in the seven treatments. Subjects had no prior experience with the experiment and were recruited from the undergraduate population of the university.[11]

Five sessions were conducted for each treatment. On average, 18 subjects participated in a session, with half of them randomly assigned to the role of a Sender and the other half to the role

[11] We conducted 25 sessions (456 subjects for the treatments *Direct Declaratives, Direct Directives, Mediated Declaratives, Mediated Directives*, and *Mediated Direct Mechanism*) via face-to-face laboratory mode with z-Tree in Spring 2018 and 10 sessions (182 subjects for the treatments *Nonconforming Mediated Declaratives* and *Mediated Declaratives with Excess Accuracy*) via real-time online mode with oTree in June 2022. We also conducted one additional session for the treatment *Direct Declaratives* via real-time online mode with oTree in June 2022 for data replication purpose and confirmed that all aspects of the observed behavior in the session were qualitatively consistent with the average behavior observed in the sessions conducted via face-to-face mode with z-Tree. The data is available upon request.

Table 12: Experimental Treatments

	Direct Talk	Mediated Talk
Declaratives	*Direct Declaratives*	*Mediated Declaratives*
Directives	*Direct Directives*	*Mediated Directives*
Direct Mechanism	N/A	*Mediated Direct Mechanism*
Nonconforming Language	N/A	*Nonconforming Mediated Declaratives*
Accuracy Control	N/A	*Mediated Declaratives with Excess Accuracy*

of a Receiver.[12] Roles remained fixed throughout a session. Subjects in a session were *randomly matched* to form groups of two with one sender and one receiver in each of the 60 rounds of the game.

In each session of the experiment conducted via the face-to-face mode, upon arrival in the lab, subjects were instructed to sit at separate computer terminals. In case of the real-time online experiment, upon arrival at the designated Zoom meeting, subjects were instructed to turn on their videos and stay in a quiet place with strong internet access. It was strictly required for subjects to turn on their videos during the entire course of the experiment. Depending on the mode, each received either a hard or an electronic copy of the experimental instructions. The instructions were read aloud using slide illustrations as an aid. A comprehension quiz and a practice round followed.

Subjects were told that there would be two equally likely situations, situation s and situation t, and their rewards in the two situations differed according to Table 1, which was shown on their decision screens. At the beginning of each round, the computer randomly selected a situation. The sender privately learned the selected situation. In the mediated-direct-mechanism treatment and the declaratives treatments, the sender chose one of two messages, "s" or "t," to send to the receiver. In the directives treatments, the choices were messages "L" and "R."

In the direct-talk treatments, the sender's chosen message was always transmitted to the paired receiver as sent. In the *Mediated Declaratives* and *Mediated Directives* treatments, message "t/R" chosen by the sender was always transmitted to the receiver as sent, while for message "s/L" messages "s/L" or "t/R" were received with equal probability.[13] In *Nonconforming Mediated Declara-*

[12] Of the 35 sessions, there were 18 with 20 subjects, 8 with 18 subjects, 5 with 16 subjects, 3 with 14 subjects, and 1 with 12 subjects.

[13] We use the notation "t/R" as a convenient shorthand for either message "t" or message "R", depending on

tives message "s" was transmitted as sent and message "t" was received as either "s" or "t" with equal probability. In the *Mediated Direct Mechanism* treatment, "t" was always transmitted as "L," and "s" was transmitted as "L" or "R" with equal probability. Subjects were informed of these mediation rules.

After receiving a message, the receiver chose one of three actions, L, C, or R. Rewards for the round were then determined based on the randomly selected situation and the receiver's action. Feedback on the situation, the sender's message, the receiver's action, and the subject's reward was provided at the end of each round. For the mediated-talk treatments, the feedback included both the message sent by the sender and the message that was transmitted to the receiver.

We randomly selected two rounds for payments. The average reward a subject earned in the two selected rounds was converted into Hong Kong Dollars at a fixed and known exchange rate of HK$1 per reward point. We also provided a show-up fee of HK$30 for those who participated in the laboratory sessions and a show-up fee of HK$40 for those who participated in the real-time online sessions. Subjects on average earned HK$125 (≈ US$16) by participating in a session that lasted 1.6 hours. The final earnings were paid in cash for all experiments we conducted in the lab, and for all sessions we conducted via the real-time online mode, they were paid via the HKUST Autopay System to the bank account each participant provided to the Student Information System (SIS).

5 Findings

We begin our report of results in Section 5.1 by comparing average terminal behavior across all seven treatments, aggregated over sessions, focusing on the role played by language. In Section 5.2 we examine heterogeneity in individual behavior and ask, for each of the seven treatments and aggregated over sessions, which fraction of individuals fit a level-k classification. Next, in Section 5.3, informed by when theory predicts that mediation makes a difference and by our findings about mediation with a non-conforming language, we focus on the difference in behavior between mediated talk with conforming languages and direct talk, aggregating over sessions as well as treatments.

5.1 The impact of language – average behavior by treatment

5.1.1 Language use by senders across treatments

Table 13 summarizes sender behavior in all seven treatments over the terminal ten rounds and contrasts observed with predicted behavior. Each entry in a panel indicates the frequency of the message sent (e.g., "t") given the state observed by the sender (e.g., s).

treatment, and also when we pool treatments; the notation "s/L" is used in the same manner.

	“s”	“t”
s	**82%**	18%
t	**86%**	14%

Direct Declaratives

	“L”	“R”
s	**90%**	10%
t	**90%**	10%

Direct Directives

	“s/L”	“t/R”
s	**86%**	14%
t	**88%**	12%

Pooled

	“s/L”	“t/R”
s	**100%**	0%
t	**100%**	0%

Predicted

(a) Direct Talk

	“s”	“t”
s	**99%**	1%
t	22%	**78%**

Mediated Declaratives

	“L”	“R”
s	**99%**	1%
t	26%	**74%**

Mediated Directives

	“s”	“t”
s	**97%**	3%
t	27%	**73%**

Mediated Direct Mechanism

	“s/L”	“t/R”
s	**99%**	1%
t	25%	**75%**

Pooled

	“s/L”	“t/R”
s	**100%**	0%
t	0%	**100%**

Predicted

(b) Mediated Talk – Conforming Languages

	“s”	“t”
s	**81%**	19%
t	**93%**	7%

Nonconforming Mediated Declaratives

	“s”	“t”
s	**100%**	0%
t	**100%**	0%

Predicted

(c) Mediated Talk – Nonconforming Language

	“s”	“t”
s	**84%**	16%
t	**75%**	25%

Mediated Declaratives with Excess Accuracy

	“s”	“t”
s	**100%**	0%
t	**100%**	0%

Predicted

(d) Mediated Talk – Conforming Language with Excess Accuracy

Table 13: Sender Behavior by Treatment – Last Ten Rounds

The two leftmost panels in Table 13(a) show terminal sender behavior in the two direct talk treatments. The key finding is that *in both of the two direct talk treatments modal language use by senders conforms with the prediction of the theory:* average sender behavior is consistent with a pooling strategy that favors the message predicted by theory (message “*s*” in *Direct Declaratives* and message "L" in *Direct Directives*).

The three leftmost panels in Table 13(b) report terminal sender behavior in the three mediated-talk treatments with conforming languages. Two characteristics of language use by senders in these treatments are worth noting: (i) *In each of the three mediated talk treatments with conforming languages, modal language use by senders agrees with the prediction of the theory:* average sender behavior is consistent with senders using a separating strategy at least 75% of the time. (ii) *In all three mediated-talk treatments with a conforming language there is a systematic departure of sender behavior from the prediction of the theory:* type t senders fail to send their separating message at least 22% of the time.

The left panel in Table 13(c) reports terminal sender behavior in the last ten rounds of *Nonconforming Mediated Declaratives. Consistent with the level-k analysis, observed modal sender behavior in mediated talk with a nonconforming language is to pool on message "s". This behavior is contrary to the equilibrium-refinement prediction, which selects the unique and efficient separating equilibrium. It also is in sharp contrast to sender behavior in mediated talk with conforming languages.* The set of equilibrium outcomes with mediation does not depend on whether the language is conforming or nonconforming. In addition, in either case iterative deletion of dominated strategies selects separation. And yet, modal observed behavior is pooling with a nonconforming language in contrast to separation with conforming languages.

The left panel in Table 13(d) reports terminal sender behavior in the last ten rounds of *Mediated Declaratives with Excess Accuracy.* In this treatment we return to conforming languages, but raise the accuracy of the mediator from $p = 0.5$ to $p = 0.9$, thus rendering separation no longer incentive compatible. Indeed, *consistent with the level-k analysis, modal sender behavior is pooling on message "s".* This suggests that for mediation to be effective, it is not enough for the noisy channel to have the appropriate structure (here, one of the messages is subject to noise and the language is conforming) – it must also have the appropriate transition probabilities (here, $p = 0.5$ rather than $p = 0.9$).

5.1.2 Language use by receivers across treatments

Table 14 summarizes receiver behavior in all seven treatments over the terminal ten rounds and contrasts observed with predicted behavior. Each entry in a panel indicates the frequency of the action taken (e.g., L) given the message observed by the receiver (e.g., "t").

The panels in the top row of Table 14(a) show terminal receiver behavior in the two direct talk treatments. The principal finding is that *in each of the two direct talk treatments modal language use by receivers for the on-path messages "s" and "L", the messages that are predicted to be received exclusively, conforms with the prediction of the theory:* the modal response to messages "s" and

	L	C	R
“s”	13%	14%	**73%**
“t”	10%	**30%**	60%

Direct
Declaratives

	L	C	R
“L”	12%	14%	**74%**
“R”	8%	**31%**	61%

Direct
Directives

	L	C	R
“s/L”	13%	13%	**74%**
“t/R”	9%	**30%**	61%

Pooled

	L	C	R
“s/L”	0%	0%	**100%**
“t/R”	0%	**100%**	0%

Predicted

(a) Direct Talk

	L	C	R
“s”	**66%**	1%	33%
“t”	5%	23%	**72%**

Mediated
Declaratives

	L	C	R
“L”	**70%**	3%	27%
“R”	1%	29%	**70%**

Mediated
Directives

	L	C	R
“L”	**68%**	0%	32%
“R”	2%	18%	**80%**

Mediated
Direct Mechanism

	L	C	R
“s/L”	**68%**	1%	31%
“t/R”	3%	23%	**74%**

Pooled

	L	C	R
“s/L”	**100%**	0%	0%
“t/R”	0%	0%	**100%**

Predicted

(b) Mediated Talk - Conforming Languages

	L	C	R
“s”	10%	6%	**84%**
“t”	0%	**81%**	19%

Nonconforming
Mediated Declaratives

	L	C	R
“s”	0%	0%	**100%**
“t”	0%	**100%**	0%

Predicted

(d) Mediated Talk – Nonconforming Language

	L	C	R
“s”	13%	12%	**75%**
“t”	6%	27%	**67%**

Mediated Declaratives
with Excess Accuracy

	L	C	R
“s”	0%	0%	**100%**
“t”	0%	0%	**100%**

Predicted

(d) Mediated Talk – Conforming Language with Excess Accuracy

Table 14: Receiver Behavior by Treatment – Last Ten Rounds

“L” is the pooling action R.[14]

[14] For the off-path messages “t” and “R”, the messages that theory predicts will not be sent and received, observed behavior departs from the level-k prediction. This is in line with our earlier remark that the level-k prediction for

The panels in the top row of Table 14(b) report terminal receiver behavior in the three mediated talk treatments with conforming languages. Two characteristics of language use by receivers are worth noting: (i) *In each of the three mediated talk treatments with conforming languages, modal language use by receivers conforms with the prediction of the theory:* average behavior is consistent with receivers using a separating strategy at least 66% of the time. (ii) *In all three treatments there are common systematic departures from the prediction of the theory:* following messages "s" and "L" there is high incidence of action R being taken and following messages "t" and "R" receivers frequently take action C.

The left panel in Table 14(c) reports receiver behavior in the last ten rounds of *Nonconforming Mediated Declaratives. Consistent with the level-k analysis, in this treatment with a nonconforming language the modal receiver action following the on-path message "s" is the pooling action R.*[15] This behavior is in sharp contrast to receiver behavior with conforming languages and counter to the equilibrium-refinement prediction. Recall that iterative deletion of dominated strategies, Pareto efficiency, the strict equilibrium requirement, and persistence all, independently, select separation in the mediated talk treatments with $p = 1/2$, regardless of the choice of language. This suggests that even in games in which multiple equilibrium selection criteria agree on a unique prediction that is invariant to the choice of language, (i) the choice of language can have a dramatic effect on behavior, and (ii) this effect can be accounted for with a level-k analysis rooted in language.

The left panel in Table 14(d) reports receiver behavior in the last ten rounds of *Mediated Declaratives with Excess Accuracy. Consistent with the level-k analysis, the modal receiver action following both messages "s" and "t" is the pooling action R.* This confirms that for mediation to be effective, both the manner and the rate of message garbling are important.

5.2 Language-anchored level-k classification of individual subjects

For each of our seven treatments, the language-anchored level-k analysis makes a prediction that is the same for levels 1 and above (and in some cases for all levels, including level 0). As we have seen, in each case observed modal behavior aligns with this prediction. At the same time, there are substantial departures from the theory. To get a clearer picture of the nature of these departures at the individual level, we classify subjects according to whether their behavior is best described as level-0, level-k with $k \geq 1$, or resists classification.

Table 15 reports the level-k classifications for both senders and receivers in all treatments. To obtain the classifications, we calculate for each subject the frequencies of observed choices that are

responses to off-path messages is somewhat arbitrary. It is worth noting that (i) the off-path messages are sent and received infrequently and (ii) that the modal observed receiver response to the off-path messages, R, is an equilibrium best reply for the receiver, as is C.

[15] In this treatment, the modal receiver response to the off-path message "t" *does* conform with level-k prediction and is consistent with an equilibrium for this game.

Table 15: Level-k Classifications

	Senders				Receivers			
Treatment	L_0	$L_{k\geq1}$	$L_{k\geq0}$	Uncl.	L_0	$L_{k\geq1}$	$L_{k\geq0}$	Uncl.
Direct Declaratives	0.12	0.74	–	0.14	0.00	0.54	–	0.46
Direct Directives	0.04	0.90	–	0.06	0.00	0.45	–	0.55
Mediated Declaratives	–	–	0.85	0.15	–	–	0.59	0.41
Mediated Directives	–	–	0.83	0.17	–	–	0.62	0.38
Mediated Direct Mechanism	–	–	0.79	0.21	–	–	0.64	0.36
Nonconforming Mediated Declaratives	0.00	0.92	–	0.08	–	–	0.94	0.06
Mediated Declaratives with Excess Accuracy	0.00	0.85	–	0.15	0.00	0.42	–	0.56

consistent with a given level for that subject's role, using data from all rounds. If the frequency for some level is no less than 70% and there is no other level with a higher frequency, the subject is classified as belonging to that level.[16] Otherwise, the subject is left unclassified. For treatments in which the L_0 prediction differs from the $L_{k\geq1}$ prediction, we report both frequencies; otherwise, we report the $L_{k\geq0}$ frequency.

We find that in all seven treatments there is a majority of sender subjects who get classified as $L_{k\geq1}$ or $L_{k\geq0}$. These are the classifications that match our theory predictions. For receiver subjects the same is true in five out of seven treatments. With one exception, the frequencies of $L_{k\geq1}$ and $L_{k\geq0}$ are higher for senders than for receivers. Few, if any, subjects are classified as L_0, consistent with the notion that the level zero type is only a mental construct, the model used by the lowest level "real" type.

Overall, while this shows that there is a general tendency to conform with the level-k prediction also at the individual level, there is substantial heterogeneity, and the conforming tendency is less pronounced for receiver subjects. It deserves emphasis that the level-k analysis correctly predicts the shift in modal observed behavior also at the individual level when switching from conforming to non-conforming languages.

5.3 The impact of mediation with conforming languages

Our findings regarding language use suggest that for mediation to be effective (i) the mediation rule has to be sufficiently inaccurate (e.g., $p = 1/2$ rather than $p = 9/10$) and (ii) the language must conform with the mechanism. With this in mind, in this section we compare the data from mediated talk with conforming languages and $p = 1/2$, aggregated over all sessions of the corresponding

[16] Having more than one level with a frequency above 70% is possible because observations may be consistent with multiple strategies. In direct declaratives, for example, the L_0 and $L_{k\geq1}$ strategies of the sender are different, but differ only in the behavior prescribed for type t.

treatments, with the data from direct talk, aggregated over all sessions of the two direct-talk treatments.

To see whether we are justified in aggregating the data across treatments, we conducted Mann-Whitney tests for differences in behavior between the two direct talk treatments and pairwise tests for differences in behavior for the three mediated-talk treatments with conforming languages. For senders, one of the differences, for type s in direct declaratives versus direct directives, is marginally significant ($p = 0.0575$). In the remaining seven of eight tests the differences are not significant ($p \geq 0.2317$). For receivers, in all eight tests the differences are not significant ($p \geq 0.4206$). We focus on terminal behavior (last 10 rounds).

5.3.1 The impact of mediation with conforming languages on sender behavior

The two rightmost panels of Tables 13(a) and of 13(b) summarize sender behavior in the direct-talk treatments and the mediated-talk treatments with conforming languages over the last 10 rounds. The main takeaway from Table 13 is that, as predicted, there is more sender separation with language-conforming mediated talk than with direct talk, and modal message use tends to conform with the theoretical prediction of the language-anchored level-k analysis in both instances.

Type t senders are the ones predicted to behave differently under direct talk and language-conforming mediation. Consistent with that prediction, type t-senders in language-conforming mediated talk separate by sending "t/R" significantly more often than do type t-senders in the direct-talk treatments ($p < 0.001$, Mann-Whitney test). Furthermore, in the last 10 rounds of the language-conforming mediated-talk treatments, type-t senders send message "t/R" significantly more often than message "s/L" ($p < 0.001$, Wilcoxon signed-rank test), whereas in the last 10 rounds of the direct-talk treatments, type-t senders send message "s/L" significantly more often than message "t/R" ($p = 0.003$, Wilcoxon signed-rank test).

This separation of type t senders under language-conforming mediated talk is not nearly as complete as theory would predict. Contrary to the prediction from theory, type t senders send message "s/L" 25% of the time in the last 10 rounds. *This failure to fully separate is puzzling: for a sender who believes receivers to be rational, every strategy that prescribes message "s/L" for type t is weakly dominated. This follows from the observation that for the receiver every strategy that responds with an action other than R to message "t/R" is strictly dominated.* Against a rational receiver, type-t senders can guarantee their maximal payoff by sending message "t/R". Why would they not send that message?

One possible explanation for this departure from predicted sender behavior under language-conforming mediated talk is that some senders believe that receivers misperceive the game as direct talk, in which case responding with actions other than R to message "t/R" is no longer strictly dominated for receivers. A second explanation is that some senders themselves misperceive the game as direct talk: under direct talk, we do expect type-t senders to send message "s/L".

To summarize, the majority of senders pools on message "s/L" under direct talk and separates under mediated talk. Under mediated talk, a sizable minority of type t sender fails to send the predicted "t/R" messages; their behavior matches predicted behavior under direct talk.

5.3.2 The impact of language-conforming mediation on receiver behavior

The panels in the second row of Table 14(a) and the second row of Table 14(b) summarize receiver behavior in the direct-talk treatments and the mediated-talk treatments with conforming languages over the last 10 rounds. Like for senders, modal behavior gravitates toward pooling with direct talk and toward separation with mediated talk.

With language-conforming mediated talk, both messages are predicted to be observed by the receiver; they are both "on path." With direct talk the only on-path message is "s/L". Thus, when we examine differences in receiver behavior, message "s/L" is of primary interest. In the last 10 rounds of the direct-talk treatments, conditional on the on-path message "s/L" the frequency of the pooling action R is 74%, significantly higher than the frequency of R in the mediated-talk treatments ($p < 0.001$, Mann-Whitney test). Furthermore, it is observed significantly more often than the second most frequent action C in the direct-talk treatments ($p < 0.001$, Wilcoxon signed-rank test). In contrast, in the last 10 rounds with language-conforming mediated talk modal receiver behavior is most consistent with separation: conditional on message "s/L" the frequency of the separating action L is 68%, significantly higher than the frequency of L in the direct-talk treatments ($p < 0.001$, Mann-Whitney test). It is also observed significantly more often than the second most frequent action R in the mediated-talk treatments ($p = 0.0232$, Wilcoxon signed-rank test). Modal receiver behavior matches the level-k prediction under direct talk for on-path messages (see footnote 14) and under language-conforming mediated talk for both messages.

Under direct talk, following the on-path message "s/L", not all of the receiver responses are the pooling action R. Some of that might be accounted for by receivers believing in over-communication by senders.[17] We also see significantly more actions R in response to off-path messages "t/R" than predicted by the level-k analysis. This, however, may simply be consequence of an arbitrary choice of how to treat receiver responses to off-path messages in the level-k analysis (see footnote 14). Both the predicted action C and the more frequently observed action R are equilibrium responses to message "t/R" in equilibria in which both sender types pool on the predicted message "s/L".

There are two noteworthy systematic departures from the level-k prediction under language-conforming mediated talk in the last 10 rounds. First, the frequency of action R conditional

[17] Crawford, Costa-Gomes and Iriberri [25] explain how to obtain over-communication and excessive credulity in information-transmission games without a preference for truth-telling.

on message "s/L" is 31%. Second, the frequency of action C conditional on message "t/R" is 23%. In both case, theory says that the frequency should be 0%. The latter departure from the theoretical prediction is of particular interest because under mediated talk any strategy that responds to message "t/R" with an action other than R is strictly dominated for the receiver. Such strategies cannot be justified by *any* receiver belief about the sender's strategy, and therefore fail to be rationalizable (Bernheim [6]; Pearce [49]).[18]

Both of these departures from the level-k prediction are consistent with a fraction of receivers misperceiving language-conforming mediation as direct talk. Taking action R following message "s/L" would be consistent with predicted behavior under direct talk. Taking action C following message "t/R" would be a best reply under direct talk to modal observed behavior under language-conforming mediation (here it is important to note that receiver subjects have access to information about their own history of messages sent to them in addition to messages they received). Thus, like for senders, it is possible to rationalize departures from the theory predictions under language-conforming mediation by having a proportion of receivers who treat language-conforming mediation as direct talk.[19]

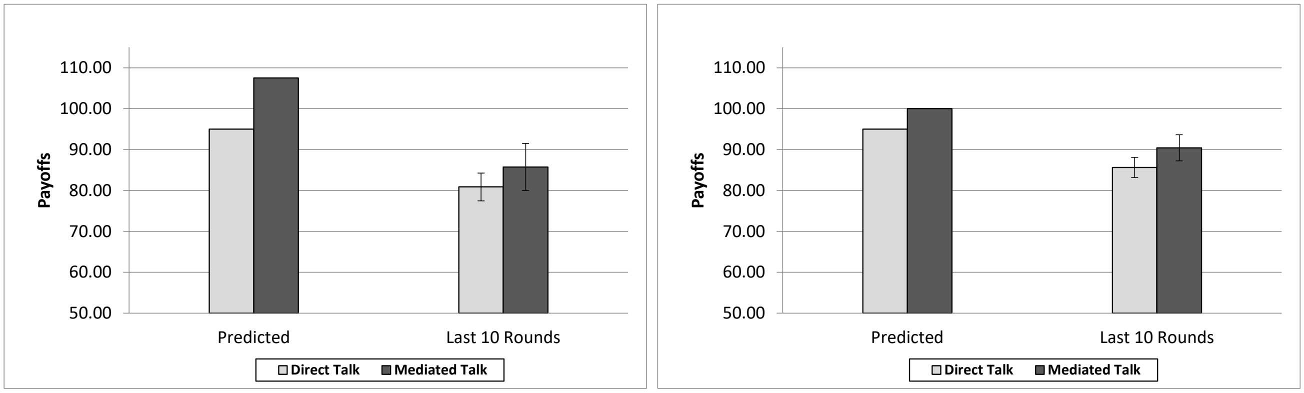
Overall, we find that, as for senders, receiver behavior under direct talk gravitates toward pooling, whereas receiver behavior under language-conforming mediation is more consistent with separation. We also again find systematic departures from this central tendency, which might be explained by a fraction of subjects treating mediated talk as direct talk.

5.3.3 The impact of language-conforming mediation on payoffs

Figure 1 compares payoffs in language-conforming mediated talk and payoffs in direct-talk over the last 10 rounds with predicted payoffs. Observed payoffs for both senders and receivers fall short of predicted payoffs in both treatments. Receivers, in particular, do not achieve even the pooling payoff with either direct or language-conforming mediated talk, despite the fact that they could guarantee that payoff by simply ignoring messages.

[18] This also rules out Quantal Response Equilibrium (QRE) (McKelvey and Palfrey [45]) as offering a systematic account of observed receiver behavior. While QRE makes it possible for action C to be observed in response to message "t/R", such behavior can only be observed as a consequence of error, given that it is strictly dominated. In addition, any departure of sender behavior from using a separating strategy makes it less attractive for the receiver to respond with action C, rather than action R, to message "t/R". In short, errors in the sender strategy should discourage the receiver from taking action C in response to message "t/R".

[19] One might object that the rationalizability puzzle – with mediated talk, there is no sender-strategy for which it would be optimal to take action C in response to message "t/R", provided the receiver fully internalizes the mediation rule – disappears if we think of our setup as a repeated game: perhaps taking action C in response to message "t/R" is part of punishment behavior. This appears unlikely. First, our random matching environment limits the scope for repeated-game effects. Second, the observed behavior is stable throughout, including terminal play, when there is no or little future that could be affected by punishment. Third, it is not clear for what the receiver would punish the sender. Under mediated talk the sender has no incentive to deceive, and therefore the receiver needs not deter deception.



(a) Senders' Average Payoffs

(b) Receivers' Average Payoffs

Figure 1: Average Payoffs in Direct Talk vs. Language-Conforming Mediation

Comparing payoffs across treatments, however, we observe higher payoffs for senders as well as receivers with language-conforming mediation than under direct talk. In the last 10 rounds under language-conforming mediation, senders and receivers receive, respectively, a payoff of 85.75 and 90.44, significantly higher than their respective average payoffs of 80.87 and 85.62 under direct talk ($p < 0.05$ Mann-Whitney tests). Figure 2 expresses these payoff gains from language-conforming mediation as a percentage of the theoretically predicted gain. Senders realize approximately 40% and receivers close to 100% of the possible payoff improvement due to replacing direct talk by language-conforming mediation.

Given that realized payoffs are substantially lower than predicted payoffs under both direct talk and language-conforming mediated talk, one may ask how well subjects do against observed play. In the last 10 rounds, senders in the direct-talk treatments could have attained a payoff of 83.68 by best-responding to the empirical distribution of receiver behavior. Their average realized payoff of 80.87 is 97% of this payoff from best-responding. With language-conforming mediation, the sender payoff from best-responding is 86.49, while the average realized payoff is 85.75, which is 99% of the payoff from best-responding. Receivers in the direct-talk treatments could have attained a payoff of 95.49 by best-responding to the empirical distribution of sender behavior. Their average realized payoff of 85.62 is 90% of this payoff from best-responding. With language-conforming mediation, the receiver payoff from best-responding is 96.66, while the average realized payoff is 90.44, which is 94% of the payoff from best-responding. With the exception of receivers under direct talk, subjects do only slightly worse than by responding to the realized distribution of play.

In sum, while subjects do not have strong incentives to adjust behavior and payoffs are lower than predicted in either treatment, there is a payoff advantage from language-conforming mediation over direct talk: subjects realize a substantial fraction (approximately 40% for senders and nearly 100% for receivers) of the payoff gain from language-conforming mediation that theory predicts.

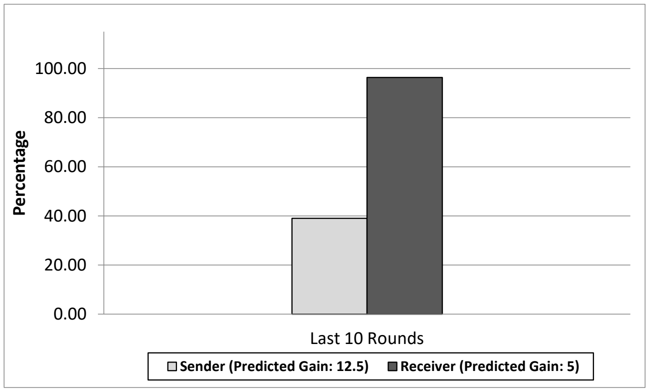


Figure 2: Observed Payoff Gain as a Percentage of the Predicted Payoff Gain

5.3.4 The impact of language-conforming mediation on behavior over time

Figure 3 compares sender behavior over time under direct talk with that under language-conforming mediation. The figure presents the 5-round moving averages of the frequencies of messages "s/L" and "t/R" conditional on type s and type t, aggregated over all direct-talk sessions in panel (a) and aggregated over all language-conforming mediation sessions in panel (b).[20]

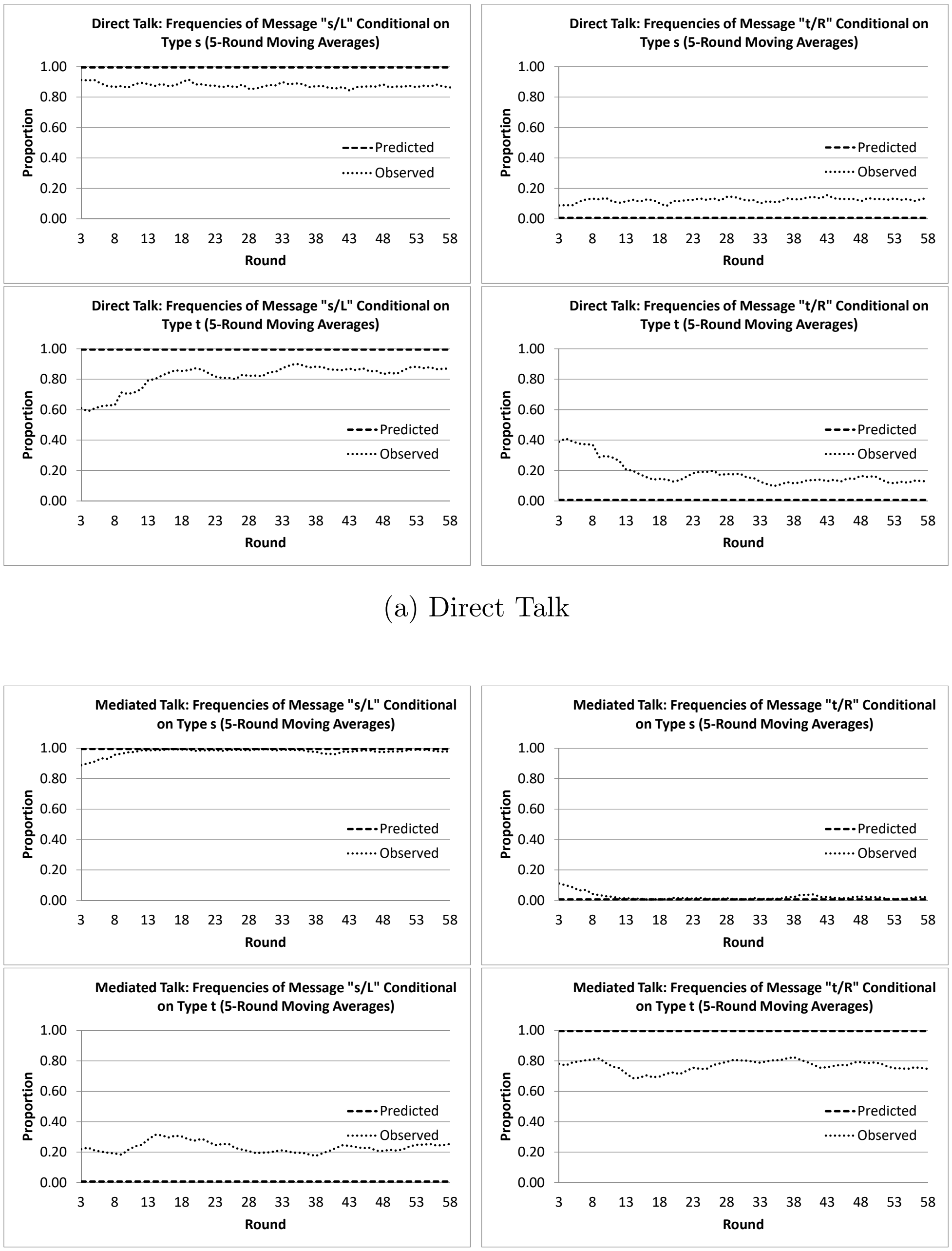
Under direct talk, type s senders send message "s/L" with a high and stable frequency throughout. Type t senders likewise send message "s/L" with higher frequency than message "t/R" from the beginning. While initially there is a substantial fraction of type t senders sending message "t/R", thus exhibiting over-communication, that fraction gradually diminishes until it stabilizes at a level below 20%. In the final 30 rounds both types s and t consistently send message "s/L" with a frequency above 80%, with no trace of over-communication.

Under language-conforming mediation, after a brief adjustment, nearly 100% of types s consistently send message "s/L", and just above 70% of types t send message "t/R" throughout. Thus, while the overall tendency is toward separation right from the outset, senders also display persistent under-communication. As we noted earlier, this under-communication can be rationalized as a fraction of senders misinterpreting language-conforming mediation as direct talk.

Figure 4 reports average receiver behavior over time. The figure presents the 5-round moving averages of the frequencies of actions L, C and R conditional on message "s/L" and message "t/R" under direct talk and language-conforming mediation.

With direct talk, conditional on the on-path message "s/L" the frequency of action R gradually rises, starting at around 40% and eventually reaching about 70%. Initially the bulk of the remaining responses to "s/L" can be attributed to credulous receiver behavior, where the receiver

[20] The moving average for round n is calculated by averaging the frequencies in rounds $n-2$, $n-1$, n, $n+1$, and $n+2$. The data points accordingly start at Round 3 and end at Round 58.



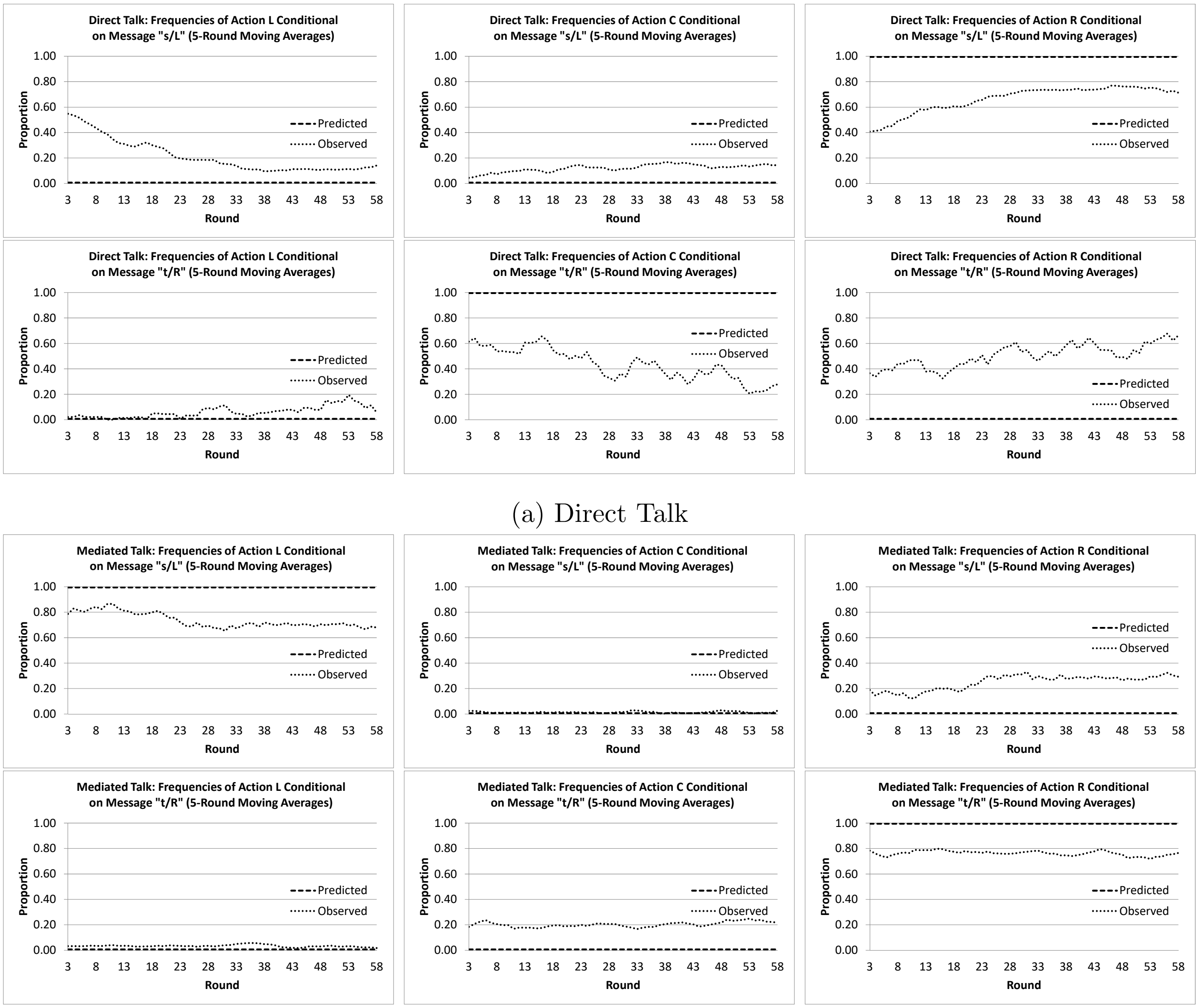
(a) Direct Talk

(b) Mediated Talk - Conforming Languages

Figure 3: Trends of Frequencies of Messages Conditional on Types (5-Round Moving Averages)

best-responds to assuming that the sender either honestly declares her type or is honest about her most preferred action. Toward the end, there is no pattern in responses to the on-path message "s/L" that are not R. This is consistent with initial over-communication disappearing with experience.

Under mediated talk, there is a stable tendency toward separation throughout. After message "s/L", initially roughly 80% of the actions taken are action L; that frequency drops slightly over time, stabilizing at around 70% during the last 30 rounds. After message "t/R", the frequency



(a) Direct Talk

(b) Mediated Talk - Conforming Languages

Figure 4: Trends of Frequencies of Actions Conditional on Messages (5-Round Moving Averages)

of action R is consistently above 70% over the entire 60 rounds. Departures from separation are likewise stable and in two directions. Following message "s/L" in excess of 20% of actions taken are action R during the last 30 rounds and following message "t/R" about 20% of the actions taken are action C over the entire 60 rounds. We noted earlier that these departures from the theoretical prediction can be rationalized by having a fraction of receivers misinterpret language-conforming mediation as direct talk

5.3.5 The impact of language-conforming mediation on individual behavior

In order to better understand how individual behavior responds to the introduction of language-conforming mediation, we regress individual choice variables on treatment and experience variables, using all-round subject-level data from direct talk and language-conforming mediation.

For senders, we regress $\mathbb{I}\{m_{i,\tau} = \text{“}t/R\text{”}\}$, an indicator variable that takes the value 1 (and zero otherwise) if sender i sends message “t/R” in period τ, on (interactions of) the following three variables: $\mathbb{I}\{\theta_{i,\tau} = t\}$, an indicator variable that takes the value 1 if the sender i's type is t in period τ; $\mathbb{I}\{M\}$, an indicator variable that takes the value 1 in the mediated talk treatment; and, $\mathbb{I}\{m_{i,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = R\}$, a product of indicator variables that equals 1 exactly when the sender used message “t/R” in the previous period and that message resulted in type t's favorite action R. For notational convenience, and slightly abusing notation, we use t, M, and P to denote these three variables.

The regression equation (for the linear probability model) is:

$$\begin{aligned}&\mathbb{I}\{m_{i,\tau} = \text{“}t/R\text{”}\} = \\ &\alpha_0 + \alpha_1 t + \alpha_2 M + \alpha_3 P + \alpha_4(t \times M) + \alpha_5(t \times P) + \alpha_6(M \times P) + \alpha_7(t \times M \times P) + \epsilon\end{aligned}$$

Theory predicts that message “t/R” is only ever sent by type t and only under mediated talk. This suggests that α_4, the treatment effect, is positive. In addition, it is possible that history matters. Specifically, a positive experience with message “t/R”, i.e., $P = \mathbb{I}\{m_{i,\tau-1} = \text{“}t/R\text{”}\} \times \mathbb{I}\{a_{i,\tau-1} = R\} = 1$ might be expected to increase the probability of sending message “t/R” if the type is t. In that case, we expect α_5 to be positive as well. Finally, if there are sender subjects with a truth-telling preference, it is possible that α_1 is positive. The signs of the remaining coefficients are not easily pinned down by theoretical, experiential, or behavioral considerations. We expect those coefficient to be not significantly different from zero.

Table 16 reports the estimated coefficients for three different model specifications, a fixed-effects linear probability model, a random-effects linear probability model, and a random-effects probit model. Since M is constant over rounds for any given subject, the treatment indicator is collinear with the subject-level fixed effects, and thus M is omitted in the fixed-effects estimation. To address the omitted M and to promote comparability between the linear probability and the probit model, we report both the fixed- and random-effects estimates for the linear model. The estimates from the two methods are not too different. For the probit model, we report the implied average marginal effects rather than the estimated coefficients themselves.

The coefficient of $t \times M$ is positive and highly significant in all three specifications, confirming the treatment effect predicted by theory. Likewise, the coefficient of $t \times P$ is positive and highly significant in all three specifications, indicating that experience matters. The coefficient of t is small and not statistically significant, suggesting that truth-telling preferences do not play a noticeable role in explaining sender behavior. There is no straightforward explanation for the negative and highly significant coefficient on M in two of the specifications; given that in one specification the estimated

Table 16: Linear Probability and Probit Models: Senders

	(1)	(2)	(3)
Constant	0.080^{***}	0.118^{***}	–
	(0.013)	(0.014)	–
t	0.037	0.036	0.031
	(0.021)	(0.021)	(0.018)
M	–	-0.080^{***}	-0.118^{***}
	–	(0.015)	(0.020)
P	−0.038	0.038	−0.005
	(0.042)	(0.037)	(0.022)
$t \times M$	0.631^{***}	0.630^{***}	0.440^{***}
	(0.048)	(0.047)	(0.025)
$t \times P$	0.325^{***}	0.324^{***}	0.139^{***}
	(0.062)	(0.058)	(0.028)
$M \times P$	−0.053	-0.090^{*}	−0.049
	(0.045)	(0.039)	(0.031)
$t \times M \times P$	−0.056	−0.054	0.053
	(0.071)	(0.067)	(0.033)
No. of Observations	13,452	13,452	13,452

Note: The dependent variable is an indicator for message "t/R". Column (1) reports the coefficients from estimating the fixed-effects linear probability model. Since M is constant over rounds for any given subject, the treatment indicator is collinear with the subject-level fixed effects, and thus M is omitted in the fixed-effects estimation. Column (2) reports the coefficients from estimating the random-effects linear probability model. Column (3) reports the *average marginal effects* from estimating the random-effects probit model. Since the reported numbers are marginal effects, no constant term is included. Robust standard errors clustered at the session level are in parentheses. *** indicates significance at 0.1% level, ** significance at 1% level, and * significance at 5% level.

coefficient is small, one should perhaps not read too much into this. One possible explanation might be that there are subjects who use message "s/L" under mediated talk to abdicate responsibility for which message is received.

For receivers, we regress $\mathbb{I}\{a_{i,\tau} = L\}$, an indicator variable that takes the value 1 if receiver i takes action "L" in period τ, on (interactions of) the following three variables: $\mathbb{I}\{\hat{m}_{i,\tau} = \text{“}s/L\text{”}\}$, an indicator variable that takes the value 1 if the receiver observes message "s/L" in period τ; $\mathbb{I}\{M\}$, an indicator variable that takes the value 1 in the mediated talk treatment; and, $\mathbb{I}\{\tilde{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{i,\tau-1} = s\}$, a product of indicator variables that equals 1 exactly when the receiver notices that the sender sent message "s/L" in the previous period and the sender's type was s.[21]

[21] We also consider a specification in which the experience variable $\mathbb{I}\{\tilde{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{i,\tau-1} = s\}$ is replaced by $\mathbb{I}\{\hat{m}_{i,\tau-1} = \text{“}s/L\text{”}\} \times \mathbb{I}\{\theta_{i,\tau-1} = s\}$, thus substituting received messages $\hat{m}_{i,\tau-1}$ for sent messages $\tilde{m}_{i,\tau-1}$. The result is virtually identical.

For notational convenience, and slightly abusing notation, we use S, M, and E to denote these three variables.

The regression equation (for the linear probability model) is:

$$\mathbb{I}\{a_{i,\tau} = L\} =$$
$$\beta_0 + \beta_1 S + \beta_2 M + \beta_3 E + \beta_4 (S \times M) + \beta_5 (S \times E) + \beta_6 (M \times E) + \beta_7 (S \times M \times E) + \epsilon$$

Table 17: Linear Probability and Probit Models: Receivers

	(1)	(2)	(3)
Constant	0.044**	0.058**	–
	(0.016)	(0.018)	–
S	0.148***	0.148***	0.129***
	(0.034)	(0.034)	(0.036)
M	–	−0.023	−0.059*
	–	(0.020)	(0.029)
E	0.002	0.003	0.012
	(0.017)	(0.016)	(0.017)
$S \times M$	0.552***	0.551***	0.354***
	(0.067)	(0.067)	(0.033)
$S \times E$	0.011	0.011	−0.004
	(0.023)	(0.022)	(0.015)
$M \times E$	−0.013	−0.014	−0.035
	(0.018)	(0.017)	(0.023)
$S \times M \times E$	0.017	0.018	0.030
	(0.029)	(0.028)	(0.025)
No. of Observations	13,452	13,452	13,452

Note: The dependent variable is an indicator for action L. Column (1) reports the coefficients from estimating the fixed-effects linear probability model. Since M is constant over rounds for any given subject, the treatment indicator is collinear with the subject-level fixed effects, and thus M is omitted in the fixed-effects estimation. Column (2) reports the coefficients from estimating the random-effects linear probability model. Column (3) reports the *average marginal effects* from estimating the random-effects probit model. Since the reported numbers are marginal effects, no constant term is included. Robust standard errors clustered at the session level are in parentheses. *** indicates significance at 0.1% level, ** significance at 1% level, and * significance at 5% level.

Theory predicts that action L is only ever taken following message "s/L" and only under mediated talk. This suggests strongly that β_4 is positive. If the receiver observes that message "s/L" was sent by a type s sender in the past, this might be expected to increase the probability of taking action L if the present message is "s/L". This suggests that β_5 is positive if experience matters. Finally, if there are credulous receiver subjects, it is possible that β_1 is positive. The signs of the remaining coefficients are not easily pinned down by theoretical, experiential, or behavioral considerations. We expect those coefficient to be not significantly different from zero.

Table 17 reports the estimated coefficients for three different model specifications, a fixed-effects linear probability model, a random-effects linear probability model, and a random-effects probit model. Since M is constant over rounds for any given subject, the treatment indicator is collinear with the subject-level fixed effects, and thus M is omitted in the fixed-effects estimation. To address the omitted M and to promote comparability between the linear probability and the probit model, we report both the fixed- and random-effects estimates for the linear model. The estimates from the two methods are not too different. For the probit model, we report the implied average marginal effects rather than the estimated coefficients themselves.

The coefficient of $S \times M$ is positive and highly significant in all three specifications, confirming the treatment effect predicted by theory. The coefficient of $S \times E$ is small and not statistically significant in all three specifications, suggesting that experience, as modeled, does not play an important role for receivers. The coefficient of S is highly statistically significant, indicating credulous behavior on part of at least some receiver subjects. Receiver credulity would also help explain why there is a persistent non-trivial fraction of receiver subjects responding with action C to message "t/R" – action C is the unique best reply to message "t/R" for a receiver who takes that message at face value and interprets mediated talk as direct talk.

In order to examine receiver credulity more closely we regress $\mathbb{I}\{a_{i,\tau} = C\}$, an indicator variable that takes the value 1 (and zero otherwise) if receiver i takes action C in period τ, on (interactions of) the following two variables: $\mathbb{I}\{\hat{m}_{i,\tau} = \text{“}t/R\text{”}\}$, an indicator variable that takes the value 1 if receiver i observes message "t/R" in period τ and $\mathbb{I}\{M\}$, an indicator variable that takes the value 1 in the mediated talk treatment. For notational convenience use T, and M to denote these two variables.

The regression equation (for the linear probability) model is :

$$\mathbb{I}\{a_{i,\tau} = C\} = \gamma_0 + \gamma_1 T + \gamma_2 M + \gamma_3 (T \times M) + \epsilon$$

Receiver credulity amounts to having $\gamma_1 > 0$. At the same time, if the choice of action C is influenced by how strongly receivers perceive the game as direct talk and it is the case that a significant fraction of receivers understand the mediated-talk game correctly, we should see fewer C responses in response to message "t/R" under mediated talk. Hence, we would expect that $\gamma_3 < 0$.

Table 18 reports the estimated coefficients for three different model specifications, a fixed-effects linear probability model, a random-effects linear probability model, and a random-effects probit model. We find that the estimated value of γ_1 is positive and significant in all three specifications, consistent with receiver credulity playing a role. The estimated value of γ_3 is negative and significant

Table 18: Linear Probability and Probit Models: Receivers (Additional Analysis)

	(1)	(2)	(3)
Constant	0.058***	0.119***	–
	(0.012)	(0.010)	–
T	0.326***	0.325***	0.209***
	(0.033)	(0.033)	(0.020)
M	–	−0.103	−0.230***
	–	(0.012)	(0.021)
$T \times M$	−0.140***	−0.139***	0.059*
	(0.042)	(0.042)	(0.027)
No. of Observations	13,680	13,680	13,680

Note: The dependent variable is an indicator for action C. Column (1) reports the coefficients from estimating the fixed-effects linear probability model. Since M is constant over rounds for any given subject, the treatment indicator is collinear with the subject-level fixed effects, and thus M is omitted in the fixed-effects estimation. Column (2) reports the coefficients from estimating the random-effects linear probability model. Column (3) reports the *average marginal effects* from estimating the random-effects probit model. Since the reported numbers are marginal effects, no constant term is included. Robust standard errors clustered at the session level are in parentheses. *** indicates significance at 0.1% level, ** significance at 1% level, and * significance at 5% level.

in two of our specifications; in the third specification mediation has a direct dampening effect on the frequency of action C, independent of the message received. We interpret this as strong evidence for credulity playing a role in determining receiver choices and some support for the claim that mediation reduces credulous responses to receiving message "t/R".

6 Discussion

Comparing mediated with direct talk necessitates paying close attention to the language employed: (i) While it is implicit in direct talk that sender and receiver languages coincide, this is less natural with mediation. Framing the language in terms of declaratives (as is the case in most of the extant experimental research on direct talk), for example, would have the mediator systematically transform truthful sender messages into lies sent to the receiver. (ii) Unlike with direct talk, equilibria under mediated talk need not be exchangeable – it is no longer the case that equilibrium outcomes are preserved under arbitrary permutations of messages. As a result, equilibria under mediation are closely intertwined with the language in which the mediation scheme is framed. (iii) If mediation is by design, language is a feature of this design, and the above gives us reason to believe that this design feature may matter.

Consistent with these observations, our communication design exercise establishes a key role for the language employed, both for prediction and performance. We find that Crawford's [24] language-anchored level-k analysis is a good predictor of modal behavior, even when it is in sharp conflict with the common prediction from a whole array of alternative selection criteria. We also demonstrate that mediation improves on direct communication, as long as it is language conforming, but fails otherwise.

Under mediation, switching from a conforming to a nonconforming language induces a pronounced change in the language-anchored level-k prediction. The level-k analysis predicts separation when the language conforms with the mediation scheme, and pooling otherwise. This exactly matches the change in modal behavior that we see in the data. This is remarkable because there are many good reasons for why one might expect behavior to be invariant to this modification of the language: (i) the choice of language has no impact on the set of possible equilibrium outcomes; (ii) iterative deletion of dominated strategies uniquely selects separation irrespective of the language; (iii) whatever the language, separation supports the unique Pareto efficient outcome; and, (iv) irrespective of the language, the separating equilibrium is the unique strict equilibrium, constitutes the unique persistent set, and also forms the unique minimal curb set.

Other language variations help putting the language-anchored level-k analysis to the test. Importantly, we exploit the fact that the language-anchored level-k analysis goes beyond predicting outcomes with direct talk. Instead, it makes sharp predictions about which messages are used, and how they are used. We demonstrate that the level-k analysis correctly predicts modal message use by senders under direct talk, even as the language is varied.

The possibility of players communicating with directives instead of declaratives (in both direct and mediated talk) implies that the language-anchored level-k analysis cannot simply be taken off the shelf. Crawford [24] anchors his analysis in senders being truthful (as do Cai and Wang [14], and Wang, Spezio, and Camerer [51]). Truthfulness, however, has no meaning with directives (Sobel [50]). We find that anchoring the analysis in "forthrightness" correctly predicts modal behavior in environments in which truthfulness has no meaning. This extends the reach of the language-anchored level-k analysis and lends further support to it.

Access to a language in which either truthful behavior is possible or desirable also has a bearing on whether there is a role for lying aversion (Abeler et al. [1]) in our setting. It makes sense to speak of "lying" when the sender's language consists of declaratives (Sobel [50]). In contrast, one cannot lie with directives. If lying aversion is a significant feature of preferences, one might expect more sender separation when the sender's language is framed in terms of declaratives rather than directives. We do not find such an effect. It is conceivable that lying aversion plays a role in explaining the difference in observed behavior between mediation with conforming declaratives and mediation with a non-conforming declaratives. In the former case, the unique efficient equilibrium is consistent with senders behaving truthfully, whereas in the latter case it requires senders to lie.

While this is suggestive, note that the difference in behavior we do observe is already accounted for through the language-anchored level-k analysis.

Varying the language lets us compare direct with indirect mechanisms, which has broader implications for mechanism design (compare Masatlioglu, Taylor, and Uler [44]). One of our conforming-language treatments considers a direct mechanism in which separation is supported by an equilibrium in which senders are truthful and receivers obedient (Myerson [46]). We contrast this with mechanisms in which either truthfulness or obedience have no meaning, as well as with a mechanism in which the equilibrium that supports separation requires senders to be anti-truthful. We find that the direct mechanism performs as well as or better than the alternatives. It performs strictly better than a mechanism that requires anti-truthful behavior to sustain separation.

The language-anchored level-k analysis captures modal behavior well. There are, however, substantial and sometimes stable departures from the theoretical prediction. Confirming results from prior work on communication games, we find over-communication by senders in the initial rounds of direct talk; this over-communication vanishes with experience. In contrast, there is stable under-communication by senders under mediated talk with conforming languages. Receivers over-interpret messages under both direct talk and mediated talk with conforming languages, i.e., with non-negligible frequency they take actions that would only be optimal if they had more precise information. This is particularly striking under mediated communication with conforming languages, where a sizable fraction of receiver subjects use strictly dominated strategies. Departures from theory under mediated talk with conforming languages are largely consistent with some of the subjects perceiving mediated communication as direct communication.

The challenge for future work will be to see whether the lessons learned from our deliberately chosen simple setting extend to other communication environments. Specifically, it would be nice to know whether there is a robust mediation scheme that facilitates communication for a broad range of incentives and with relaxed knowledge requirements about the nature of private information.

References

[1] Abeler, Johannes, Daniele Nosenzo, and Collin Raymond [2019], "Preferences for truth-telling," *Econometrica* **87**, 1115-1153.

[2] Ambrus, Attila, Eduardo M. Azevedo, Yuichiro Kamada, and Yuki Takagi [2013], "Legislative Committees as Information Intermediaries: A Unified Theory of Committee Selection and Amendment Rules," *Journal of Economic Behavior & Organization* **94**, 103-115.

[3] Ambrus, Attila, Eduardo M. Azevedo, and Yuichiro Kamada [2013], "Hierarchical Cheap Talk," *Theoretical Economics* **8**, 233-261.

[4] Au, Pak Hung, OSub Kwon, King King Li [2022], "A Simple Experiment on Simple Bayesian Persuasion," working paper, HKUST.

[5] Basu, Kaushik, and Jörgen W. Weibull [1991], "Strategy Subsets Closed under Rational Behavior," *Economics Letters* **36**, 141-146.

[6] Bernheim, B. Douglas [1984], "Rationalizable Strategic Behavior," *Econometrica* **52**, 1007-1028.

[7] Bichler, Martin, Paul Milgrom, and Gregor Schwarz [2022], "Taming the Communication and Computation Complexity of Combinatorial Auctions: The FUEL Bid Language," *Management Science*, https://doi.org/10.1287/mnsc.2022.4465.

[8] Bichler, Martin, Jacob Goeree, Stefan Mayer, and Pasha Shabalin [2014], "Spectrum Auction Design: Simple Auctions for Complex Sales," *Telecommunications Policy* **38**, 613-622.

[9] Blume, Andreas, Oliver J. Board, and Kohei Kawamura [2007], "Noisy Talk," *Theoretical Economics* **2**, 395-440.

[10] Blume, Andreas, Douglas V. Dejong, Yong-Gwan Kim, and Geoffrey B. Sprinkle [2001], "Evolution of Communication with Partial Common Interest," *Games and Economic Behavior* **37**, 79-120.

[11] Blume, Andreas, Ernest K. Lai, and Wooyoung Lim [2020], "Strategic Information Transmission: A Survey of Experiments and Theoretical Foundations," in C. Monica Capra, Rachel Croson, Mary Rigdon and Tanya Rosenblat (eds), *Handbook of Experimental Game Theory*, Cheltenham, UK and Northampton, MA, USA: Edward Elgar Publishing.

[12] Blume, Andreas, Ernest K. Lai, and Wooyoung Lim [2019], "Eliciting Private Information with Noise: The Case of Randomized Response," *Games and Economic Behavior* **113**, 356-380.

[13] Bühler, Karl [1934], *Sprachtheorie: Die Darstellungsfunktion der Sprache,* Verlag Gustav, Jena.

[14] Cai, Hongbin, and Joseph Tao-Yi Wang [2006], “Overcommunication in Strategic Information Transmission Games,” *Games and Economic Behavior* **56**, 7-36.

[15] Calsamiglia, Caterina, Guillaume Haeringer, and Flip Klijn [2010], “Constrained School Choice: An Experimental Study,” *American Economic Review* **100**, 1860-74.

[16] Casella, Alessandra, Evan Friedman, and Manuel Perez [2020], “Mediating Conflict in the Lab,” Working Paper.

[17] Chassang, Sylvain, and Gerard Padró i Miquel [2019], “Crime, Intimidation, and Whistleblowing: A Theory of Inference from Unverifiable Reports,” *The Review of Economic Studies* **86**, 2530-2553.

[18] Chassang, Sylvain, and Christian Zehnder [2019], “Secure Survey Design in Organizations: Theory and Experiments,” NBER Working Paper.

[19] Chen, Daniel L, Martin Schonger, and Chris Wickens [2016], “oTree - An open-source platform for laboratory, online, and field experiments,” *Journal of Behavioral and Experimental Finance* **9**, 88-97.

[20] Chen, Yan, and John O. Ledyard [2010], “Mechanism Design Experiments,” in *Behavioural and Experimental Economics* 191-205, Palgrave Macmillan, London.

[21] Chen, Yan, and Tayfun Sönmez [2006], “School Choice: An Experimental Study,” *Journal of Economic theory* **127**, 202-231.

[22] Chierchia, Gennaro and Sally McConnell-Ginet [2000], *Meaning and Grammar* Cambridge, MA, MIT Press.

[23] Crawford, Vincent P. and Joel Sobel [1982], “Strategic Information Transmission,” *Econometrica* **50**, 1431-1451.

[24] Crawford, Vincent P. [2003], “Lying for Strategic Advantage: Rational and Boundedly Rational Misrepresentation of Intentions,” *American Economic Review* **93**, 133-149.

[25] Crawford, Vincent P., Miguel A. Costa-Gomes, and Nagore Iriberri [2013] “Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications,” *Journal of Economic Literature* **51**, 5-62.

[26] Dickhaut, John W., Kevin A. McCabe, and Arijit Mukherji [1995], “An Experimental Study of Strategic Information Transmission,” *Economic Theory* **6**, 389-403.

[27] Duffy, John, and Nick Feltovich [2010], "Correlated Equilibria, Good and Bad: an Experimental Study," *International Economic Review* **51(3)**, 701-721.

[28] Fischbacher, Urs [2007], "z-tree: Zurich Toolbox for Ready-made Economic Experiments," *Experimental Economics* **10**, 171-178.

[29] Forges, Françoise [1985], "Correlated Equilibria in a Class of Repeated Games with Incomplete Information," *International Journal of Game Theory* **14**, 129-149.

[30] Fréchette, Guillaume, Alessandro Lizzeri, Jacopo Perego [2022], "Rules and Commitment in Communication: An Experimental Analysis," *Econometrica* **90**, 2283-2318.

[31] Goltsman, Maria, Johannes Hörner, Gregory Pavlov, and Francesco Squintani [2009], "Mediation, Arbitration and Negotiation," *Journal of Economic Theory* **144**, 1397-1420.

[32] Gordon, Sidartha, Navin Kartik, Melody Pei-yu Lo, Wojciech Olszewski, and Joel Sobel [2021], "Effective Communication in Cheap Talk Games," Working Paper.

[33] Hakimov, Rustamdjan, and Dorothea Kübler [2021], "Experiments on Centralized School Choice and College Admissions: A Survey," *Experimental Economics* **24**, 434-488.

[34] Harris, Milton, and Artur Raviv [2010], "Control of Corporate Decisions: Shareholders vs. Management," *The Review of Financial Studies* **23**, 4115-4147.

[35] Hörner, Johannes, Massimo Morelli, and Francesco Squintani [2015], "Mediation and Peace," *The Review of Economic Studies* **82**, 1483-1501.

[36] Hurkens, Sjaak [1995], "Learning by Forgetful Players," *Games and Economic Behavior* **11**, 304-329.

[37] Ivanov, Maxim [2010], "Communication via a Strategic Mediator," *Journal of Economic Theory* **145**, 869-884.

[38] Jakobson, Roman [1960], "Linguistics and Poetics," in Thomas A. Sebeok (ed.) *Style in Language*, MA: MIT Press, 350-377.

[39] Kalai, Ehud, and Dov Samet [1984], "Persistent Equilibria in Strategic Games," *International Journal of Game Theory* **13**, 129-144.

[40] Kamenica, Emir, and Matthew Gentzkow [2011], "Bayesian Persuasion," *American Economic Review* **101**, 2590-2615.

[41] Krishna, Vijay, and John Morgan [2004], "The Art of Conversation: Eliciting Information from Experts through Multi-Stage Communication," *Journal of Economic Theory* **117**, 147-179.

[42] Laclau, Marie, Ludovic Renou, and Xavier Venel [2020], "Robust Communication on Networks," Working paper.

[43] Lewis, David [1969], *Convention. A Philosophical Study,* Harvard University Press, Cambridge, MA.

[44] Masatlioglu, Yusufcan, Sarah Taylor, and Neslihan Uler [2012], "Behavioral mechanism design: evidence from the modified first-price auctions," *Review of Economic Design* **16**, 159-173.

[45] McKelvey, Richard D. and Thomas R. Palfrey [1995], "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior* **10**, 6-38.

[46] Myerson, Roger B. [1982], "Optimal Coordination Mechanisms in Generalized Principal-Agent Problems," *Journal of Mathematical Economics* **10**, 67-81.

[47] Myerson, Roger B. [1991], *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge, MA.

[48] Nguyen, Quyen [2016], "Bayesian Persuasion: Evidence from the Laboratory," Utah State University, Working Paper.

[49] Pearce, David G. [1984], "Rationalizable Strategic Behavior and the Problem of Perfection," *Econometrica* **52**, 1029-1050.

[50] Sobel, Joel [2020], "Lying and Deception in Games," *Journal of Political Economy* **128**, 907-947.

[51] Wang, Joseph Tao-yi, Michael Spezio, and Colin F. Camerer [2010], "Pinocchio's Pupil: Using Eyetracking and Pupil Dilation to Understand Truth Telling and Deception in Sender-Receiver Games," *American Economic Review* **100**, 984-1007.

[52] Warner, Stanley L. [1965], "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias," *Journal of the American Statistical Association* **60**, 63-69.

Online Appendix

Equilibrium Analysis and additional data analysis of direct talk and mediated talk with conforming languages

A Equilibrium analysis

For the equilibrium analysis the framing of messages as directives or declaratives does not matter: given any equilibrium under one frame, there exists an outcome-equivalent equilibrium under any alternative frame, in which the message are simply renamed.

A.1 The set of equilibrium outcomes with direct talk

The receiver has seven possible types of responses to any message received: he can randomize over all three actions, L, C, and R , randomize over two, e.g. C and R, or take a single action with probability one. Any randomization with both L and C in the support is ruled out as a best response, because whenever the receiver is indifferent between L and C he strictly prefers R. In any pairwise comparison of the remaining five response types, the sender in state s strictly prefers one of them. It follows that if two messages are sent with positive probability in equilibrium and the receiver responds differently after those message, the sender in state s will have a strict preference for one of the messages, leaving the other message sent exclusively in state t. This, however, cannot constitute an equilibrium, because the message that is exclusively sent in state t identifies the state and leads to the least preferred action for the sender. Therefore, in any equilibrium either only one message is sent or the receiver's responses do not vary with the messages. In both cases, the receiver takes action R in response to any message that is sent in equilibrium. It follows that ***pooling is the only equilibrium outcome with direct communication***. There are many equilibria that support the pooling outcome. The equilibrium analysis does not discriminate among these equilibria, although it seems reasonable to expect that the framing of messages impacts behavior.

A.2 The set of equilibrium outcomes with mediated talk and $p = 1/2$

Since the framing of messages is irrelevant for the equilibrium analysis, we will take advantage of the notational convenience of conducting this analysis in the specific framework of mediated declaratives. Recall that with mediated declaratives the set of sent messages $\{s, t\}$ coincides with the set of received messages. Pooling, where the receiver responds to all messages received in equilibrium with action R, is supported by many equilibria. Suppose instead that the receiver observes both messages with positive probability in equilibrium, and responds differently to different messages. Call such an equilibrium *influential.* Since the receiver's expected posterior equals the prior, he will mix (possibly degenerately) over L and R after one message and over C and R after the other. We will refer to the former type of lottery by LR and the latter by CR.

Regarding influential equilibria, there are two cases to consider:

Case 1: The receiver uses LR after message s.

By sending message s the sender induces LR with probability one half and CR otherwise. By sending message t she induces CR with probability one. Since at least one of the two lotteries assigns positive probability to an action other than R, type s strictly prefers sending message s. For the receiver to treat messages s and t differently type t must send message t with probability greater than zero. Suppose type t sends message t with probability one. Then the receiver assigns posterior probability $\frac{2}{3}$ to type t after receiving message t. Thus action R is the unique best reply after receiving message t. Therefore we have an equilibrium in pure strategies in which sender type s sends message s, sender type t sends message t, the receiver responds to message s with action L and to message t with action R.

Suppose type t sends message t with probability $x \in (0,1)$. Then the posterior probability of type t given message t is

$$\gamma := \frac{x\frac{1}{2}}{x\frac{1}{2} + \sigma(s|s)\frac{1}{2}\frac{1}{2}} = \frac{2x}{2x+1}.$$

The posterior probability of type t given message s is

$$\delta := \frac{(1-x)\frac{1}{2}\frac{1}{2}}{(1-x)\frac{1}{2}\frac{1}{2} + \sigma(s|s)\frac{1}{2}\frac{1}{2}} = \frac{1-x}{2-x}.$$

For the receiver to be indifferent between C and R following message t would require that $120\gamma = 90\gamma + 100(1-\gamma)$, or $\gamma = \frac{10}{13}$. For this we would need $x = \frac{5}{3}$, which is impossible. For the receiver to be indifferent between L and R following message s would require that $120(1-\delta) = 90\delta + 100(1-\delta)$, or $\delta = \frac{2}{11}$. For this we need $x = \frac{7}{9}$. With $x = \frac{7}{9}$ the receiver takes action R with probability one after receiving message t. This gives the sender a choice between inducing an LR lottery that assigns probability less than one to R and inducing R with probability one. This implies that sender type t strictly prefers sending message t, which contradicts $x \in (0,1)$.

Thus for the case under consideration, the only influential equilibrium is the one in in which type s sends message s and type t sends message t, which supports the outcome we refer to as *separation.*

Case 2: The receiver uses LR after message t.

By sending message t, the sender induces LR with probability one. By sending message s the sender induces LR with probability one half and CR otherwise. Since at least one of the two lotteries assigns probability less than one to action R, type s of the sender strictly prefers sending message t. For the receiver to respond differently after the two messages, type t must send message s with probability greater than zero. Then, since only type t sends message s with positive probability the receiver's unique best reply to message s is action C, which results in sender type t receiving her lowest possible payoff. Since the case assumes that the receiver mixes between L and R following message t, type t has a strict preference for message t over message s. This implies that the two types pool on message t contradicting our assumption that we have an influential equilibrium. Thus there is no influential equilibrium in this case.

In summary, since an equilibrium that is not influential is a pooling equilibrium, we have shown that ***with mediation separation and pooling are the only two equilibrium outcomes.*** Furthermore, we have found that ***there is a unique equilibrium supporting separation***, which implies that ***for the case of separation, the equilibrium analysis pins down message use.*** For the case of pooling, the equilibrium analysis does not pin down message use.

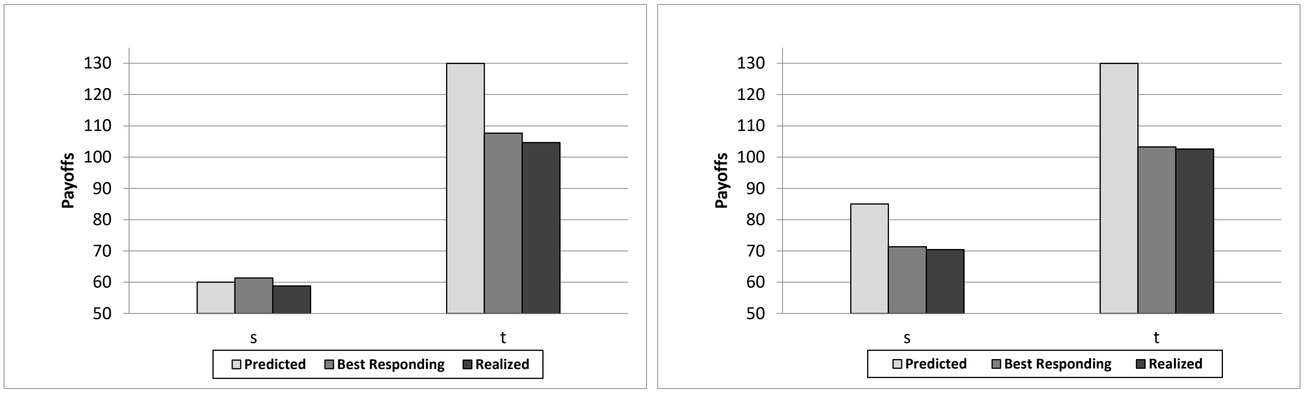
B Payoffs from best responses to empirical play

We compare realized average payoffs in the last 10 rounds with "counter-factual" payoffs that subjects would have realized had they best responded to the "empirical strategies" of opponents in the last 10 rounds.

B.1 Senders

For senders, we compute the relative frequencies of actions L, C, and R conditional on the two messages in the last 10 rounds of each session. These frequencies serve as a proxy for receiver behavior strategies, which we use to determine sender best-responses. For each session, we identify which of the two messages yields a higher expected payoff to a hypothetical sender of a given type who plays against this empirically determined behavior strategy of receiver. The payoffs thus identified, one for each type, become the sender best-responding payoffs for the session.

We aggregate these session-level best-responding payoffs across direct-talk or mediated-talk treatments and compare them with the average realized-type payoffs of senders in the last 10 rounds of the corresponding sets of treatments. Figure 5 presents the aggregate payoffs.[22] To facilitate comparison, we also include in the figure the payoffs predicted by theory.



(a) Direct Talk

(b) Mediated Talk

Figure 5: Sender Best-Responding and Realized Payoffs by Type

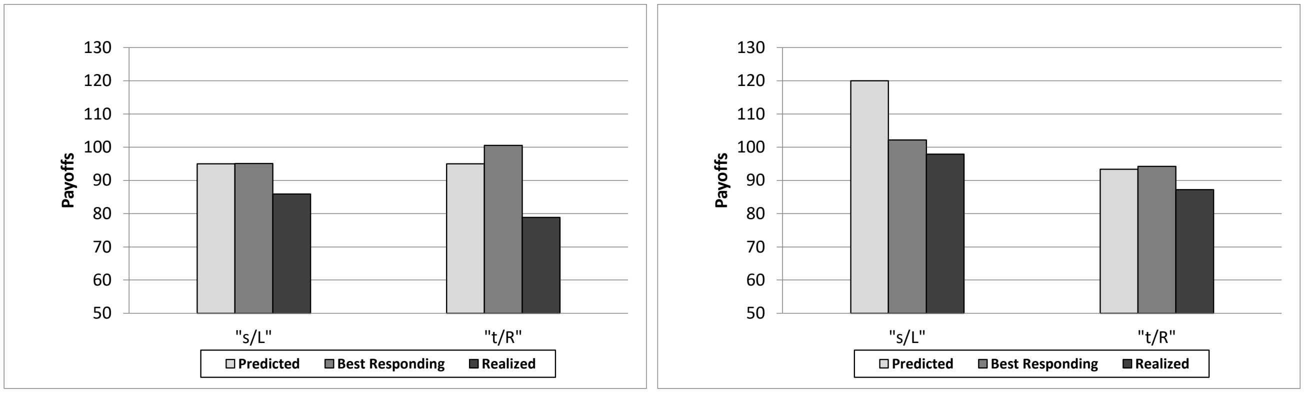
For the direct-talk treatments, the sender best-responding payoffs are 61.37 for type s and 107.67 for type t. The average realized payoffs are, for type s, 96% and, for type t, 97% of these best-responding payoffs. For the mediated-talk treatments, the sender best-responding payoffs are 71.31 for type s and 103.3 for type t. The average realized payoffs are, for both types, 99% of these best-responding payoffs.

[22] In determining best-responding payoffs for the mediated-talk treatments, we use observed frequencies of the mediated messages conditional on "s/L" being sent, which are not exactly one half but in close neighborhoods of the uniform randomization specified in the game.

B.2 Receivers

For receivers, we compute the relative frequencies of types s and t conditional on the two messages *received* in the last 10 rounds of each session. These frequencies serve as a proxy for the beliefs induced by those messages. For each session, we identify which of the three actions yields a higher expected payoff to a hypothetical receiver upon receiving a message given these empirically computed beliefs that are consistent with how senders use message. The payoffs thus identified, one for each message received, become the receiver best-responding payoffs for the session.

We aggregate these session-level best-responding payoffs across direct-talk or mediated-talk treatments and compare them with the average realized payoffs of receivers for each of the two messages received in the last 10 rounds of the corresponding sets of treatments. Figure 6 presents the aggregate payoffs. To facilitate comparison, we also include in the figure the payoffs predicted by the theory.



(a) Direct Talk

(b) Mediated Talk

Figure 6: Receiver Best-Responding and Realized Payoffs by Message Received

For the direct-talk treatments, the receiver best-responding payoffs are 95.12 for message "s/L" and 100.53 for message "t/R." The average realized payoffs are, for message "s/L," 90% and, for message "t/R," 78% of these best-responding payoffs. Note that for message "t/R," the best-responding payoff is higher than the predicted payoff of 95. Message "t/R" is used infrequently. Based on the limited observations, there are a few sessions in which it is sent more often by type s than by type t. This raises the expected payoff from the best-responding R above 95 and in some cases even makes action L the best-responding action with a payoff as high as 120.

For the mediated-talk treatments, the receiver best-responding payoffs are 102.18 for message "s/L" and 94.22 for message "t/R." The average realized payoffs are, for message "s/L," 96% and, for message "t/R," 93% of these best-responding payoffs.

C Individual treatments

In this section we disaggregate and report data separately for each individual treatment.

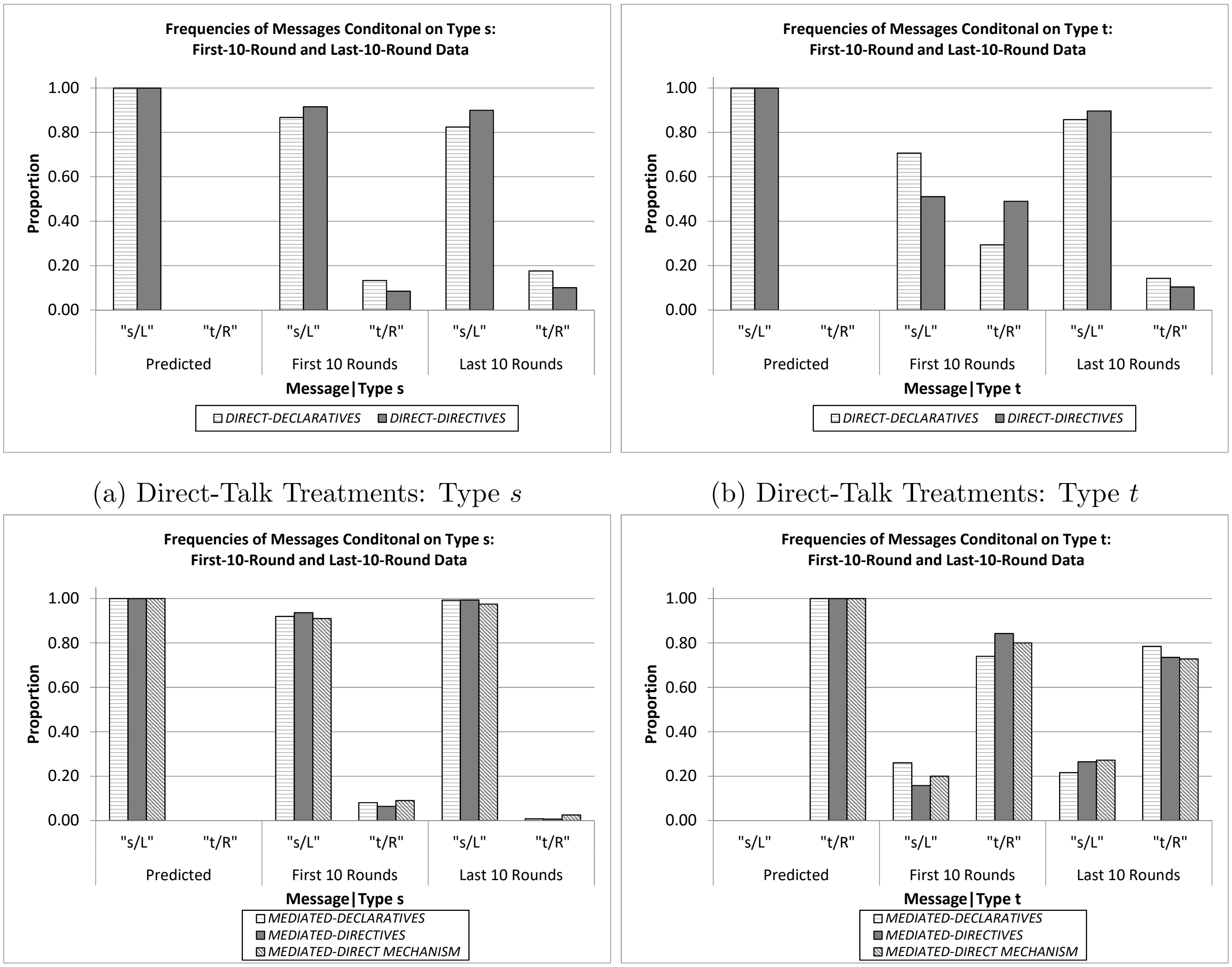
C.1 Individual treatments: senders

Figure 7 shows sender behavior in each of the five treatments, in both the first 10 rounds and the last 10 rounds, and relates observations to predictions.

The top panels report sender behavior in each of the two direct-talk treatments. There is a noticeable difference between the two treatments in the first 10 rounds, which tends to vanish in the last 10 rounds, especially for type-t behavior. Type-s senders in the first 10 rounds send message "L" 92% of the time in the *Direct-Directives* treatment, more frequently than the 87% of message "s" in the *Direct-Declaratives* treatment ($p = 0.0476$, Mann-Whitney test); in the last 10 rounds they send message "L" 90% of the time in the *Direct-Directives* treatment, also more frequently than the 82% of message "s" in the *Direct-Declaratives* treatment but with a slightly lower statistical significance ($p = 0.0575$, Mann-Whitney test). For type-t senders, in the first 10 rounds they send message "R" 49% of the time in the *Direct-Directives* treatment, more frequently than the 29% of the message "t" in the *Direct-Declaratives* treatment ($p = 0.0159$, Mann-Whitney test); in the last 10 rounds they send message "R" 10% of the time in the *Direct-Directives* treatment, less frequently than the 14% of message "t" in the *Direct-Declaratives* treatment but with no statistical significance ($p = 0.2317$, Mann-Whitney test).

The behavior of type-t senders in the initial rounds is consistent with them trying to *direct* receivers to take their favorite action R: while our sender-anchored level-k analysis, where the anchor is forthright behavior of senders at level 0, does not distinguish between the direct-directives and direct-declaratives treatments, in a receiver-anchored level-k analysis, where the anchor is credulity of receivers at level 0 of the direct-directives game, at level 0 type ts separate, sending message R, whereas in the analysis of the direct-declaratives game they pool, sending message s. Overall, given the slight difference in initial behavior and no discernible difference in terminal behavior, there appears to be no loss in pooling the data from both treatments.

The lower panels of Figure 7 report sender behavior in each of the three mediated talk treatments. In all three treatments the modal behavior is separation from the outset, with no substantial difference across treatments. In the first 10 rounds, type-s senders send message "s" 92% of the time in the *Mediated-Declaratives* treatment, send message "L" 94% of the time in the *Mediated-Directives* treatment, and send message "s" 91% of the time in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.4206$ in any pairwise comparison, Mann-Whitney tests); type-t senders send message "t" 74% of the time in the *Mediated-Declaratives* treatment, send message "R" 84% of the time in the *Mediated-Directives* treatment, and send message "t" 80% of the time in the



(a) Direct-Talk Treatments: Type s

(b) Direct-Talk Treatments: Type t

(c) Mediated-Talk Treatments: Type s

(d) Mediated-Talk Treatments: Type t

Figure 7: Senders' Behavior in Individual Treatments:
First-10-Round and Last-10-Round Data

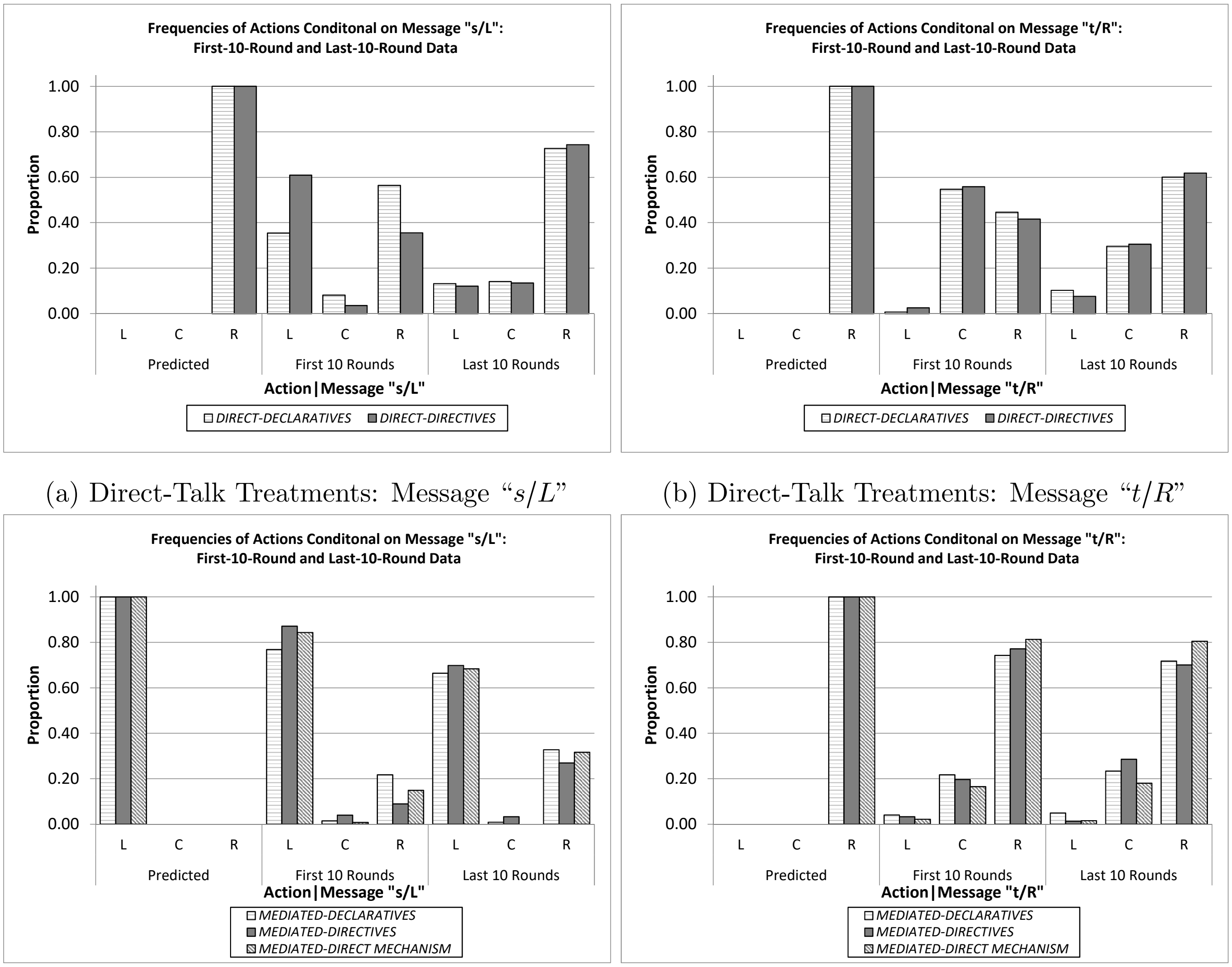
Mediated-Direct Mechanism treatment (two-sided $p \geq 0.2222$ in any pairwise comparison, Mann-Whitney tests). In the last 10 rounds, type-s senders send message "s" 99% of the time in the *Mediated-Declaratives* treatment, send message "L" 99% of the time in the *Mediated-Directives* treatment, and send message "s" 97% of the time in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.2652$ in any pairwise comparison, Mann-Whitney tests); type-t senders send message "t" 78% of the time in the *Mediated-Declaratives* treatment, send message "R" 74% of the time in the *Mediated-Directives* treatment, and send message "t" 73% of the time in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.6905$ in any pairwise comparison, Mann-Whitney tests).[23] In all three treatments separation is more pronounced for s types than for t types. Absent significant differences across treatments, it makes sense to pool the data from the three mediated-

[23] For each of the four cases, a Kruskal-Wallis test further confirms that the three frequencies have no statistical differences from one another ($p \geq 0.2894$).

talk treatments.

C.2 Individual treatments: receivers

Figure 8 reports receiver behavior in each of the five treatments, in the first 10 rounds and the last 10 rounds.



(a) Direct-Talk Treatments: Message "s/L"

(b) Direct-Talk Treatments: Message "t/R"

(c) Mediated-Talk Treatments: Message "s/L"

(d) Mediated-Talk Treatments: Message "t/R"

Figure 8: Receivers' Behavior in Individual Treatments: First-10-Round and Last-10-Round Data

The top panels show receiver behavior in each of the two direct-talk treatments. In the last 10 rounds there is almost no difference in receiver behavior between the two direct-talk treatments. Conditional on message "s/L," the frequency of action R is 73% in the *Direct-Declaratives* treatment and 74% in the *Direct-Directives* treatment; conditional on message "t/R," the frequency of action R is 60% in the *Direct-Declaratives* treatment and 62% in the *Direct-Directives* treatment ($p \geq 0.8413$, Mann-Whitney tests). In the first 10 rounds, there is also no noticeable difference in receiver behavior following message "t/R" between the two treatments, but following message "s/L" receivers respond more frequently with action L in the *Direct-Directives* treatment than in the *Direct-Declaratives* treatment. Conditional on message "t/R," the frequency of action C is 55% in the *Direct-Declaratives* treatment and 56% in the *Direct-Directives* treatment (two-sided p = 0.6905, Mann-Whitney test); conditional on message "s/L," the frequency of action L is 61% in the *Direct-Directives* treatment, significantly higher than the 35% in the *Direct-Declaratives* treatment ($p < 0.01$, Mann-Whitney test). There is some support for this kind of initial behavior in the receiver-anchored level-k analysis: level-1 receivers respond with action L to message "L" under direct directives, whereas they respond with action R to message "s" under direct declaratives. This difference between treatments is slight and vanishes over time, suggesting that there is no loss in pooling the data from the two direct-talk treatments.

The lower panels of Figure 8 show receiver behavior in each of the three mediated-talk treatments. In all three treatments the modal behavior is separation from the outset, with no significant variations across the treatments. In the first 10 rounds, conditional on message "s/L," the frequency of action L is 77% in the *Mediated-Declaratives* treatment, 87% in the *Mediated-Directives* treatment, and 84% in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.402$ in any pairwise comparison, Mann-Whitney tests); conditional on message "t/R," the frequency of action R is 74% in the *Mediated-Declaratives* treatment, 77% in the *Mediated-Directives* treatment, and 81% in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.222$ in any pairwise comparison, Mann-Whitney tests).[24]

Departures towards action R following message "s/L" and toward action C following message "t/R" are observed in terminal behavior, which are common to all three treatments. In the last 10 rounds, conditional on message "s/L," the frequency of action R is 33% in the *Mediated-Declaratives* treatment, 27% in the *Mediated-Directives* treatment, and 32% in the *Mediated-Direct Mechanism* treatment (two-sided p = 0.9166 in any pairwise comparison, Mann-Whitney tests); conditional on message "t/R," the frequency of action C is 23% in the *Mediated-Declaratives* treatment, 29% in the *Mediated-Directives* treatment, and 18% in the *Mediated-Direct Mechanism* treatment (two-sided $p \geq 0.4206$ in any pairwise comparison, Mann-Whitney tests).[25] The homogeneity of the three mediated talk treatments suggests that pooling the data from the three treatments is without loss.

C.3 Individual treatments: outcomes

Table 19 reports the outcomes for the two direct-talk treatments over the first and last 10 rounds.

[24] For each of the two cases, a Kruskal-Wallis test further confirms that the three frequencies have no statistical differences from one another ($p \geq 0.3791$).

[25] For each of the two cases, a Kruskal-Wallis test further confirms that the three frequencies have no statistical differences from one another ($p \geq 0.6126$).

While it appears that initially type s is somewhat more readily identified in the *Direct-Directives* treatment, that difference disappears over time. In the last 10 rounds the outcomes from the two treatments are very similar, where the frequencies of the pooling action R are 70% in the *Direct-Declaratives* treatment and 74% in the *Direct-Directives* treatment (two-sided $p = 0.6905$, Mann-Whitney test).

	L	C	R
s	0%	0%	**50%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	15%	7%	**30%**
t	13%	10%	**25%**

First 10 Rounds

	L	C	R
s	7%	9%	**39%**
t	6%	8%	**31%**

Last 10 Rounds

(a) Direct-Declaratives

	L	C	R
s	30%	4%	**19%**
t	15%	13%	**19%**

First 10 Rounds

	L	C	R
s	6%	7%	**36%**
t	6%	8%	**37%**

Last 10 Rounds

(b) Direct-Directives

Table 19: Direct Communication Outcomes (Treatment Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

Table 20 reports the outcomes for the three mediated-talk treatments over the first and last 10 rounds. The frequencies of outcome (s, L) in the first 10 and last 10 rounds are 17% and 18% in the *Mediated-Declaratives* treatment, commonly 20% in the *Mediated-Directives* treatment, and 17% and 16% in the *Mediated-Direct Mechanism* treatment. The frequencies of outcome (t, R) in the first 10 and last 10 rounds are 36% and 32% in the *Mediated-Declaratives* treatment, 35% and 32% in the *Mediated-Directives* treatment, and 42% and 39% in the *Mediated-Direct Mechanism* treatment. The outcomes are fairly homogeneous, both across time ($p \geq 0.4227$, Wilcoxon signed-rank tests) and across treatments (two-sided $p \geq 0.1732$ in any relevant pairwise comparison, Mann-Whitney tests).[26] In each of the six panels at least 76% of the data are consistent with separation. In line

[26] For each of the four comparisons across treatments, a Kruskal-Wallis test further confirms that the three frequencies have no statistical differences from one another ($p \geq 0.2535$).

with theory, conditional on type s, the distribution over actions is bimodal, placing substantial weight on action L, the optimal action conditional on identifying type s.

	L	C	R
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	**17%**	7%	**23%**
t	7%	9%	**37%**

First 10 Rounds

	L	C	R
s	**18%**	7%	**28%**
t	5%	10%	**32%**

Last 10 Rounds

(a) Mediated-Declaratives

	L	C	R
s	**20%**	6%	**25%**
t	5%	9%	**35%**

First 10 Rounds

	L	C	R
s	**20%**	9%	**24%**
t	4%	11%	**32%**

Last 10 Rounds

(b) Mediated-Directives

	L	C	R
s	**17%**	5%	**24%**
t	5%	7%	**42%**

First 10 Rounds

	L	C	R
s	**16%**	6%	**28%**
t	4%	7%	**39%**

Last 10 Rounds

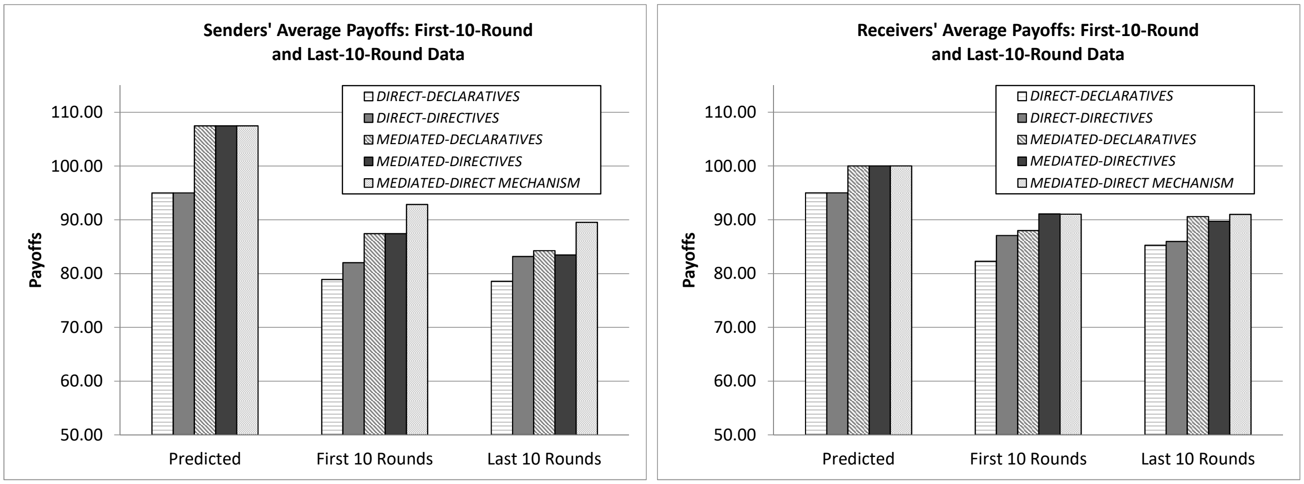
(c) Mediated-Direct Mechanism

Table 20: Mediated-Communication Outcomes (Treatment Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

C.4 Individual treatments: payoffs

Figure 9 reports the payoffs for all five treatments over the first and last 10 rounds.

Payoff differences are small. This is not that surprising given that predicted payoff differences are themselves small. It is even less surprising in light of the fact that there is substantial noise in both sender and receiver behavior. Combining sender and receiver behavior compounds that



(a) Senders' Average Payoffs

(b) Receivers' Average Payoffs

Figure 9: Average Payoffs in Direct Talk vs. Mediated Talk:
First-10-Round and Last-10-Round Data

noise when considering outcomes and payoffs. Nevertheless, regardless of whether we consider the first ten or the last 10 rounds, and for both senders and receivers, payoffs in the mediated-talk treatments are never less than payoffs in the direct talk treatments. This suggests that mediation has a positive effect on payoffs.

D Session-level data

In this section we report outcome data for each session, for the first and last 10 rounds.

D.1 Session-level data on sender behavior

Figure 10 reports sender behavior in each of the five direct-declaratives sessions. There is some heterogeneity: in Session 2 types *s* send message "*t*" at three times the rate they send it in all other sessions. Also, in the first 10 rounds there are three sessions in which types *t* send message "*t*" with considerably higher frequency than in the other two sessions; this behavior is consistent with over-communication in the first 10 rounds of those three sessions. Overall, however, there is uniformity in modal behavior. In all five sessions, in both the first 10 and the last 10 rounds, and for both types the modal message is "*s*", consistent with the level-*k* prediction.

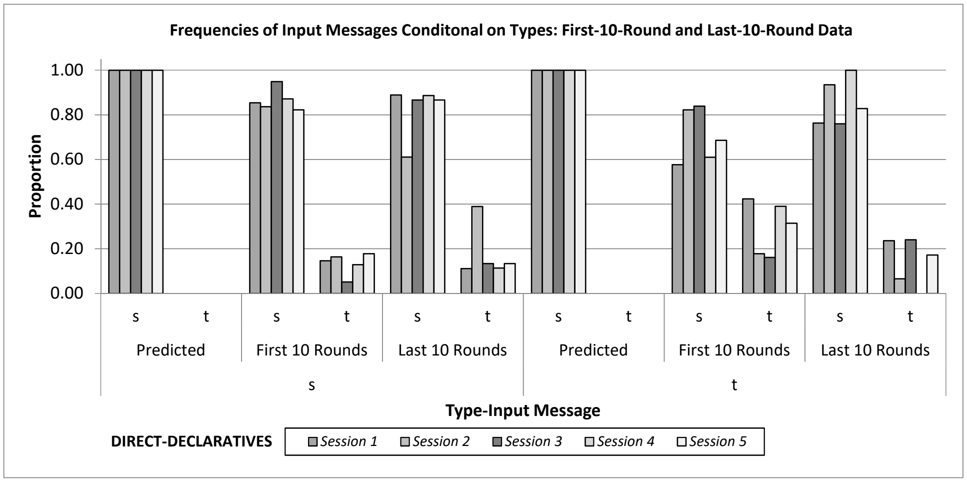


Figure 10: Senders' Behavior in Direct-Declaratives: First-10-Round and Last-10-Round Data

Figure 11 reports sender behavior in each of the five direct-directives sessions. There is considerable over-communication in the first 10 rounds: in four of the five sessions more than 40% of types t send message "R"; while this is consistent with postulated level-0 behavior, predicted behavior for all higher levels is for types t to send message "L". This over-communication disappears in the last ten rounds. There modal behavior is uniform across sessions: both types in all sessions send message "L" with at least probability 0.8. This is in line with the level-k prediction.

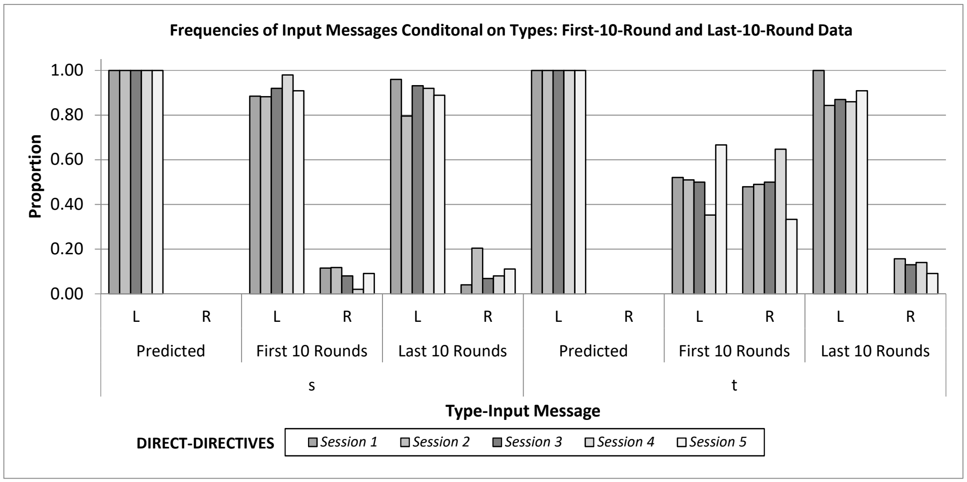


Figure 11: Senders' Behavior in Direct-Directives: First-10-Round and Last-10-Round Data

Figure 12 reports sender behavior in each of the five mediated-declaratives sessions. There is some under-communication in both the first and last 10 rounds. Even in the last 10 rounds there are two sessions in which type t senders send message "s" at least 30% of the time. Modal behavior is uniform across both the first and last ten periods and across all sessions: the majority of types s send message "s" and the majority of types t send message "t". This is the separating strategy predicted by the level-k analysis.

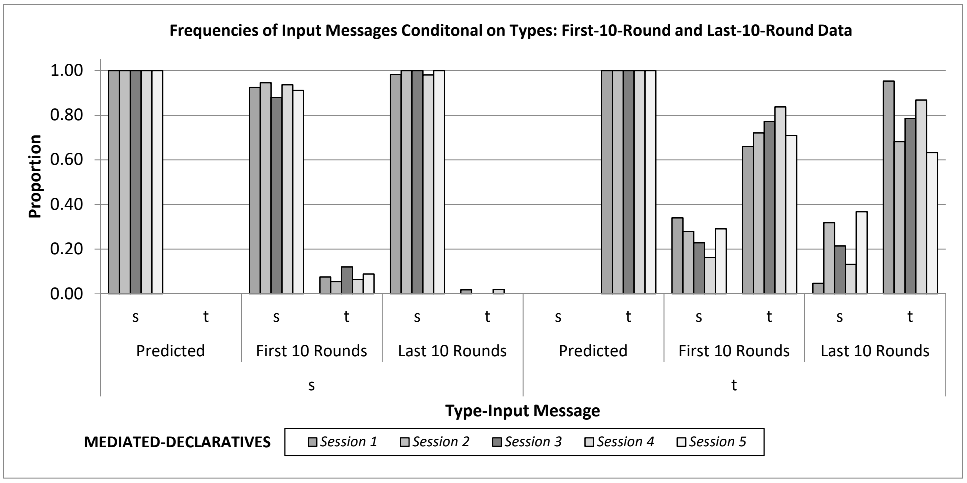


Figure 12: Senders' Behavior in Mediated-Declaratives: First-10-Round and Last-10-Round Data

Figure 13 reports sender behavior in each of the five mediated-directives sessions. There is some under-communication and heterogeneity, especially in the last 10 rounds, with one session being fully separating and another session in which types t send the two messages "L" and "R" with roughly equal probability. In the first 10 rounds modal behavior is separation in all five sessions. In the last 10 rounds modal behavior is separation in four out of five sessions. Separation is predicted by the level-k analysis.

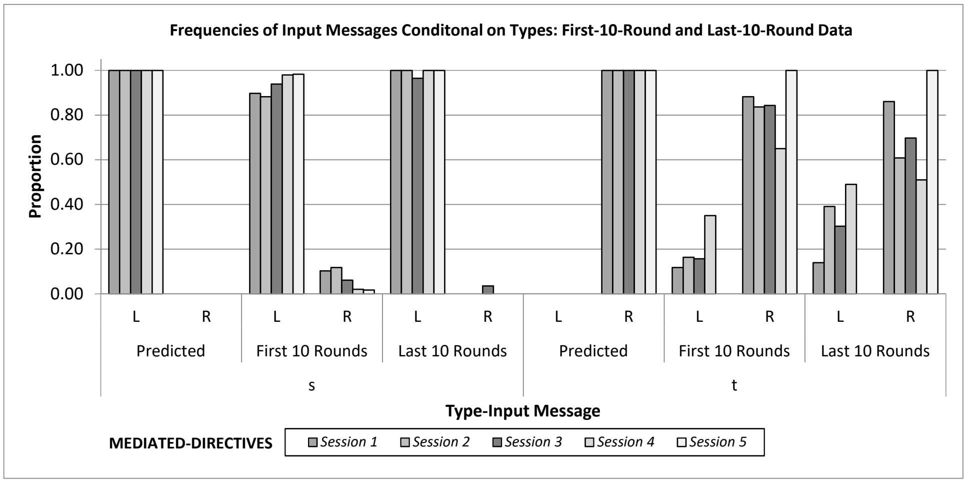


Figure 13: Senders' Behavior in Mediated-Directives: First-10-Round and Last-10-Round Data

Figure 14 reports sender behavior in each of the five mediated-direct-mechanism sessions. Again, there is some under-communication and heterogeneity, especially in the last 10 rounds, with three sessions in which types t send message "L" more than 30% of the time. In the first 10 rounds modal behavior is separation in all five sessions. In the last 10 rounds modal behavior is separation in four out of five sessions. Separation is predicted by the level-k analysis.

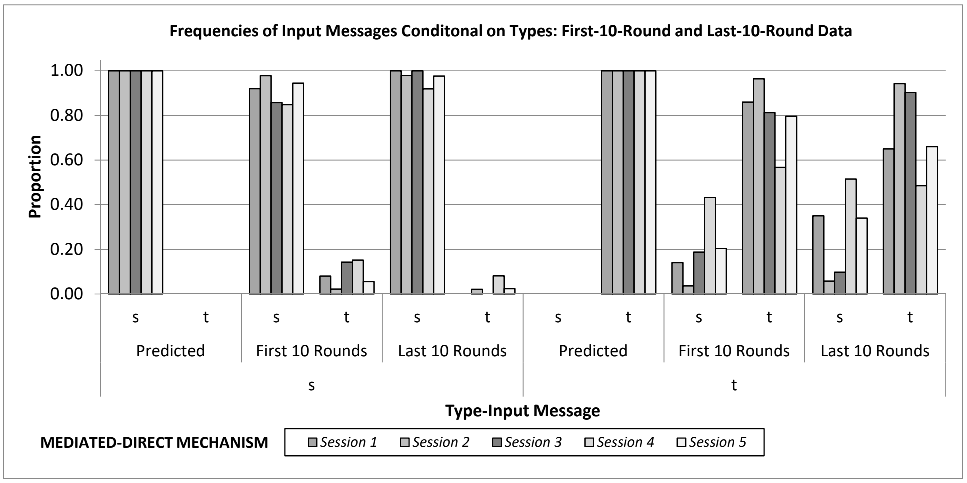


Figure 14: Senders' Behavior in Mediated-Direct Mechanism: First-10-Round and Last-10-Round Data

D.2 Session level data on receiver behavior

Figure 15 reports receiver behavior in each of the five direct-declarative sessions. There is heterogeneity in the response to message "t" both in the first and in the last 10 rounds. Message "t" should not be observed according to predicted sender behavior and is indeed observed relatively infrequently. In the first 10 rounds in three of the sessions the modal response is C. In the last 10 rounds action C is still a frequent response in two sessions, but the modal response is R in four out of five sessions. Both responses C and R to message "t" are consistent with the pooling equilibrium prediction in which only message "s" is sent, but only C is supported by the level-k analysis.

Message "s" is the only message receivers should observe according to predicted sender behavior and is also the most frequent message they do observe. In the first 10 rounds the modal response to message "s" is action R, consistent with the level-k analysis, in four of the five sessions. The other frequent response to message "s" in the first ten round is action L, which is the best response to postulated level-0 sender behavior and would be a best response with sufficient over-communication by senders. The modal response to message "s" is action R in all five sessions in the last ten periods, consistent with predicted receiver behavior.

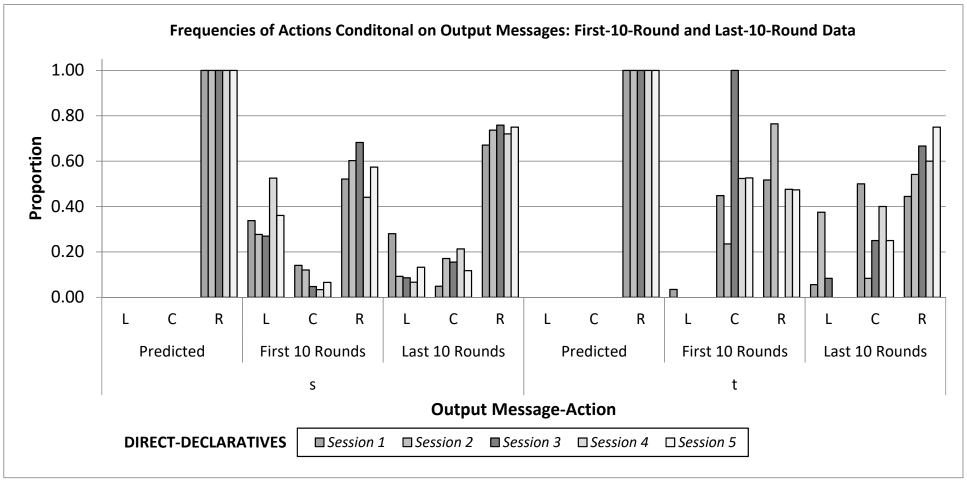


Figure 15: Receivers' Behavior in Direct-Declaratives: First-10-Round and Last-10-Round Data

Figure 16 reports receiver behavior in each of the five direct-directives sessions. There is heterogeneity in the response to message "R" both in the first and in the last 10 rounds. Message "R" should not be observed according to predicted sender behavior and is indeed observed relatively infrequently by the receiver. In the first 10 rounds in three of the sessions the receiver's modal response is C. In the last 10 rounds action C is still a frequent response in two sessions, but the modal response is R in four out of five sessions. Both responses C and R to message "R" are consistent with the pooling equilibrium prediction in which only message "L" is sent, but only C is supported by the level-k analysis. This closely resembles the pattern we observe with responses to message "t" in the direct-declaratives sessions.

Message "L" is the only message receivers should observe according to predicted sender behavior and is also the most frequent message they do observe. In the first 10 rounds the modal response to message "L" is action L in all five sessions. This is consistent with postulated level-0 behavior, but not with predicted level-k behavior for any level above 0; it would be a best reply with sufficient over-communication by senders. The modal response to message "s" is action R in all five sessions in the last ten periods, consistent with predicted receiver behavior. Thus in all five sessions there is a dramatic shift in behavior from the first 10 rounds to the last 10 rounds in the direction of the theoretical prediction.

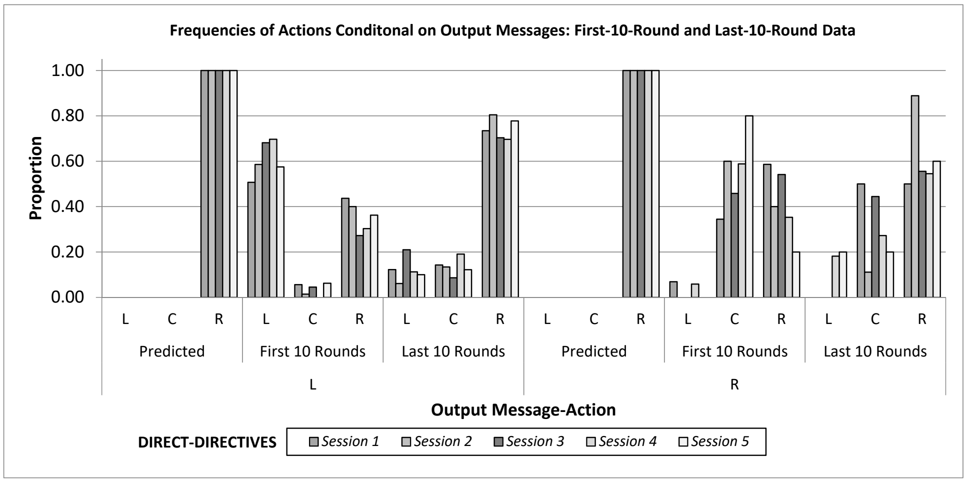


Figure 16: Receivers' Behavior in Direct-Directives: First-10-Round and Last-10-Round Data

Figure 17 reports receiver behavior in each of the five mediated-declaratives sessions. In the first 10 rounds, in four out of five sessions, the modal receiver strategy is separation, responding to message "s" with action L and to message "t" with action R. In the last 10 rounds there are two sessions in which the modal receiver strategy is separation; in two sessions the modal receiver strategy is pooling. Thus while we see separation more often than with direct talk, there is considerable heterogeneity and the tendency toward separation is more pronounced in the early rounds.

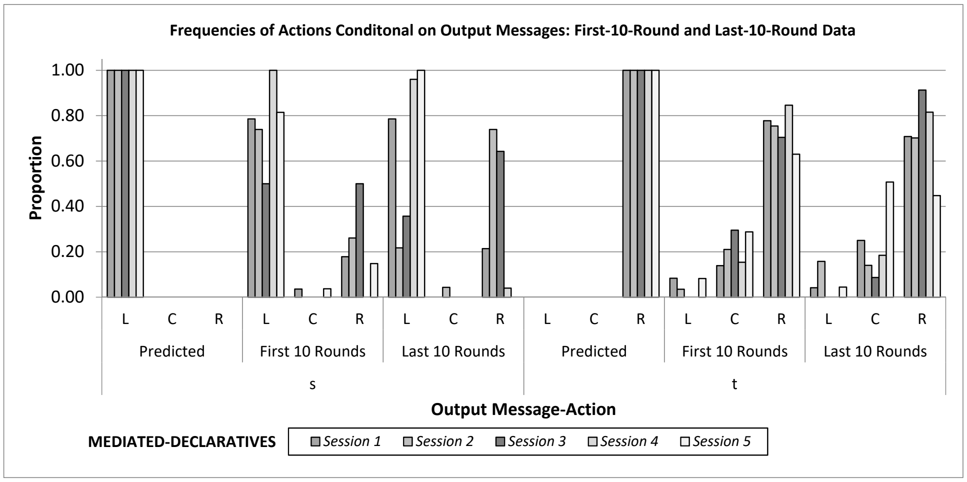


Figure 17: Receivers' Behavior in Mediated-Declaratives:
First-10-Round and Last-10-Round Data

Figure 18 reports receiver behavior in each of the five mediated-directives sessions. In the first 10 rounds, in all five sessions, the modal receiver strategy is separation, responding to message "L" with action L and to message "R" with action R. In the last 10 rounds there are four sessions in which the modal receiver strategy is separation. As with mediated declaratives, separation is more pronounced in the first ten than the last 10 rounds. Departures from separation are in the direction of taking action R following message "L" and taking action C in response to message "R". The former suggests increased pessimism about being able to extract information form message "L", while the latter suggests, increased optimism about the ability to extract information from message "R". The tendency toward separation as the modal behavior is in line with predicted behavior.

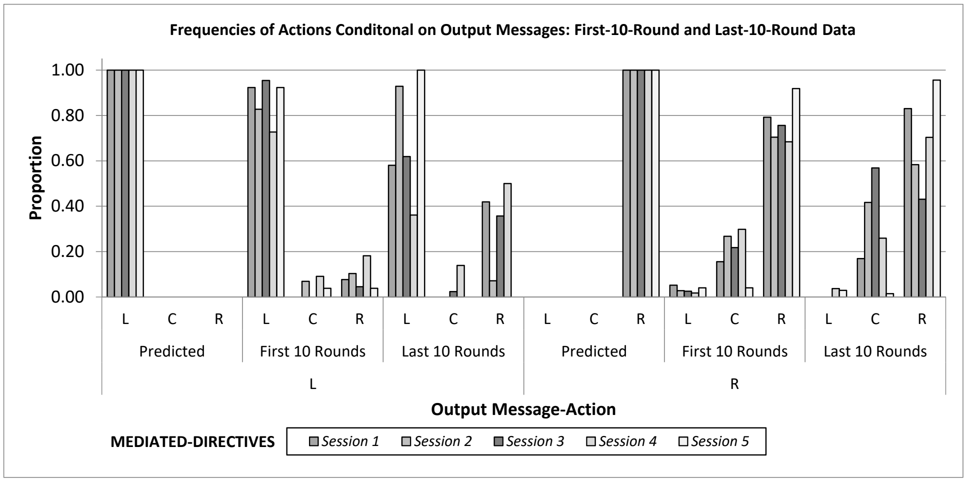


Figure 18: Receivers' Behavior in Mediated-Directives:
First-10-Round and Last-10-Round Data

Figure 19 reports receiver behavior in each of the five mediated-direct-mechanism sessions. In the first 10 rounds, in all five sessions, the modal receiver strategy is separation, responding to message "L" with action L and to message "R" with action R. In the last 10 rounds there are four sessions in which the modal receiver strategy is separation. As with mediated declaratives, separation is more pronounced in the first ten than the last 10 rounds. Departures from separation are in the direction of taking action R following message "L" and taking action C in response to message "R". The former suggests increased pessimism about being able to extract information form message "L", while the latter suggests, increased optimism about the ability to extract information from message "R". The behavior pattern in the mediated-direct-mechanism sessions closely resembles that in the mediated directives sessions. The tendency toward separation as the modal behavior is in line with predicted behavior.

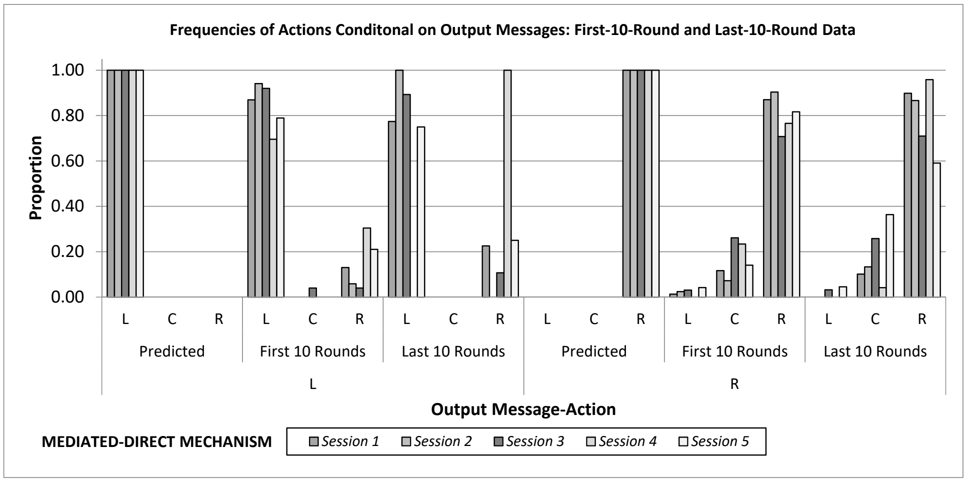


Figure 19: Receivers' Behavior in Mediated-Direct Mechanism:
First-10-Round and Last-10-Round Data

D.3 Session-level outcomes: direct declaratives

Table 21 presents the observed outcomes for each of the five direct-declaratives sessions, aggregated over the first 10 and the last 10 rounds.

In all five sessions in the last 10 rounds more than 60% of the data are consistent with pooling and in four out of five sessions more than 70% of the data are consistent with pooling. In each session there is more weight on the pooling outcome during the last 10 rounds than during the first 10 rounds. Session level data support the conclusion that in the direct-talk declaratives treatment behavior in the last 10 rounds is best described by pooling.

	L	*C*	*R*
s	0%	0%	**50%**
t	0%	0%	**50%**

Predicted

	L	*C*	*R*
s	10%	9%	**29%**
t	15%	14%	**23%**

First 10 Rounds

	L	*C*	*R*
s	10%	5%	**30%**
t	14%	8%	**33%**

Last 10 Rounds

(a) Session 1

	L	*C*	*R*
s	13%	7%	**35%**
t	10%	7%	**28%**

First 10 Rounds

	L	*C*	*R*
s	11%	7%	**36%**
t	5%	8%	**33%**

Last 10 Rounds

(b) Session 2

	L	*C*	*R*
s	13%	7%	**36%**
t	11%	7%	**26%**

First 10 Rounds

	L	*C*	*R*
s	4%	11%	**49%**
t	4%	6%	**26%**

Last 10 Rounds

(c) Session 3

	L	*C*	*R*
s	23%	5%	**21%**
t	16%	11%	**24%**

First 10 Rounds

	L	*C*	*R*
s	4%	12%	**38%**
t	3%	11%	**32%**

Last 10 Rounds

(d) Session 4

	L	*C*	*R*
s	16%	9%	**31%**
t	11%	9%	**24%**

First 10 Rounds

	L	*C*	*R*
s	6%	8%	**42%**
t	5%	6%	**33%**

Last 10 Rounds

(e) Session 5

Table 21: Communication Outcomes in Direct-Declaratives (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

D.4 Session-level outcomes: direct directives

Table 22 presents the observed outcomes for each of the five direct-directives sessions, aggregated over the first 10 and the last 10 rounds.

In all five sessions in the last 10 rounds more than 65% of the data are consistent with pooling, and the weight on pooling increases from the the first to the last 10 rounds. During the first 10 rounds there are systematic departures from pooling. In terms of our level-k analysis, the pattern of these departures from pooling is consistent with there being a mix of L_0 and of $L_{k\geq1}$ players. According tho the level-k analysis, all type-action combinations except (s, C) have positive probability. This is consistent with (s, C) being the least frequently observed type-action pair in the first 10 rounds of each of the five direct-talk directive sessions.

	L	C	R
s	0%	0%	**50%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	24%	4%	**24%**
t	14%	10%	**24%**

First 10 Rounds

	L	C	R
s	6%	6%	**38%**
t	6%	9%	**35%**

Last 10 Rounds

(a) Session 1

	L	C	R
s	28%	6%	**17%**
t	13%	13%	**23%**

First 10 Rounds

	L	C	R
s	4%	6%	**39%**
t	1%	7%	**43%**

Last 10 Rounds

(b) Session 2

	L	C	R
s	32%	3%	**21%**
t	18%	12%	**14%**

First 10 Rounds

	L	C	R
s	9%	8%	**32%**
t	10%	4%	**37%**

Last 10 Rounds

(c) Session 3

	L	C	R
s	34%	1%	**14%**
t	14%	19%	**18%**

First 10 Rounds

	L	C	R
s	8%	8%	**34%**
t	4%	12%	**34%**

Last 10 Rounds

(d) Session 4

	L	C	R
s	29%	8%	**18%**
t	17%	13%	**15%**

First 10 Rounds

	L	C	R
s	2%	7%	**36%**
t	9%	6%	**40%**

Last 10 Rounds

(e) Session 5

Table 22: Communication Outcomes in Direct-Directives (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

D.5 Session-level outcomes: mediated declaratives

Table 23 presents the observed outcomes for each of the five mediated declaratives sessions, aggregated over the first 10 and the last 10 rounds.

If we measure proximity to separation versus pooling according to whether the frequency of (s, L) outcome realizations is closer to 25% (as separation would predict) than to 0% (as pooling would predict), then there are three sessions in which the outcome in the last ten periods is closer to separation than to pooling. During the first 10 rounds four sessions would be categorized as separating according to this role of thumb. Finally, in all five sessions in the last 10 rounds (s, L) is the most frequent type-action pair in which action R is not taken.

With only three out of five sessions closer to separation than pooling, there is considerable heterogeneity across sessions in the last 10 rounds. Still, the modal outcome realization not involving action R is (s, L) in all five sessions in the last ten rounds and in four out of five sessions in the first 10 rounds, suggesting an overall tendency toward separation.

	L	*C*	*R*
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Predicted

	L	*C*	*R*
s	**17%**	8%	**28%**
t	11%	3%	**33%**

First 10 Rounds

	L	*C*	*R*
s	**23%**	10%	**24%**
t	2%	8%	**33%**

Last 10 Rounds

(a) Session 1

	L	*C*	*R*
s	**15%**	6%	**25%**
t	9%	9%	**36%**

First 10 Rounds

	L	*C*	*R*
s	**11%**	5%	**29%**
t	6%	6%	**43%**

Last 10 Rounds

(b) Session 2

	L	*C*	*R*
s	**8%**	10%	**23%**
t	5%	12%	**42%**

First 10 Rounds

	L	*C*	*R*
s	**7%**	3%	**43%**
t	2%	3%	**42%**

Last 10 Rounds

(c) Session 3

	L	*C*	*R*
s	**26%**	3%	**23%**
t	2%	8%	**38%**

First 10 Rounds

	L	*C*	*R*
s	**21%**	7%	**29%**
t	6%	7%	**30%**

Last 10 Rounds

(d) Session 4

	L	*C*	*R*
s	**21%**	7%	**17%**
t	7%	15%	**33%**

First 10 Rounds

	L	*C*	*R*
s	**27%**	9%	**15%**
t	9%	25%	**15%**

Last 10 Rounds

(e) Session 5

Table 23: Communication Outcomes in Mediated-Declaratives (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

D.6 Session-level outcomes: mediated directives

Table 24 presents the observed outcomes for each of the five mediated directives sessions, aggregated over the first 10 and the last 10 rounds.

If we measure proximity to separation versus pooling according to whether the frequency of (s, L) outcome realizations is closer to 25% than to 0% then there are four sessions in which the outcome in the last ten periods is closer to separation than to pooling. The session closer to pooling is Session 4. There are also four out of five sessions closer to separation than pooling during the first 10 rounds, and three sessions are closer to separation than pooling throughout.

With four out of five sessions closer to separation than pooling, there is heterogeneity across sessions in the last 10 rounds. Still, the modal outcome realization not involving action R is (s, L) in four out of five sessions in the last ten rounds and in all five sessions in the first 10 rounds, suggesting an overall tendency toward separation.

	L	C	R
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Predicted

	L	C	R
s	**10%**	4%	**29%**
t	8%	9%	**40%**

First 10 Rounds

	L	C	R
s	**19%**	4%	**29%**
t	1%	7%	**40%**

Last 10 Rounds

(a) Session 1

	L	C	R
s	**23%**	9%	**19%**
t	3%	12%	**34%**

First 10 Rounds

	L	C	R
s	**20%**	14%	**20%**
t	6%	16%	**24%**

Last 10 Rounds

(b) Session 2

	L	C	R
s	**19%**	7%	**23%**
t	4%	10%	**37%**

First 10 Rounds

	L	C	R
s	**19%**	16%	**22%**
t	7%	18%	**18%**

Last 10 Rounds

(c) Session 3

	L	C	R
s	**21%**	9%	**26%**
t	7%	13%	**24%**

First 10 Rounds

	L	C	R
s	**10%**	10%	**26%**
t	7%	11%	**36%**

Last 10 Rounds

(d) Session 4

	L	C	R
s	**25%**	3%	**30%**
t	2%	1%	**39%**

First 10 Rounds

	L	C	R
s	**33%**	0%	**25%**
t	1%	1%	**40%**

Last 10 Rounds

(e) Session 5

Table 24: Communication Outcomes in Mediated-Directives (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

D.7 Session-level outcomes: mediated direct mechanism

Table 25 presents the observed outcomes for each of the five mediated-talk direct mechanism sessions, aggregated over the first 10 and the last 10 rounds.

If we measure proximity to separation versus pooling according to whether the frequency of (s, L) outcome realizations is closer to 25% than to 0% then there are four sessions in which the outcome in the last 10 rounds is closer to separation than to pooling. The session (very) near to pooling is Session 4. All five sessions are closer to separation than pooling during the first 10 rounds.

With four out of five sessions closer to separation than pooling, there is heterogeneity across sessions in the last 10 rounds. Still, the modal outcome realization not involving action R is (s, L) in four out of five sessions in the last ten rounds and in all five sessions in the first 10 rounds, suggesting an overall tendency toward separation.

	L	*C*	*R*
s	**25%**	0%	**25%**
t	0%	0%	**50%**

Predicted

	L	*C*	*R*
s	**16%**	4%	**30%**
t	5%	5%	**40%**

First 10 Rounds

	L	*C*	*R*
s	**19%**	5%	**36%**
t	5%	2%	**33%**

Last 10 Rounds

(a) Session 1

	L	*C*	*R*
s	**14%**	4%	**27%**
t	4%	2%	**49%**

First 10 Rounds

	L	*C*	*R*
s	**22%**	4%	**22%**
t	3%	6%	**43%**

Last 10 Rounds

(b) Session 2

	L	*C*	*R*
s	**23%**	7%	**17%**
t	4%	13%	**36%**

First 10 Rounds

	L	*C*	*R*
s	**26%**	6%	**12%**
t	4%	12%	**40%**

Last 10 Rounds

(c) Session 3

	L	*C*	*R*
s	**17%**	6%	**24%**
t	6%	10%	**37%**

First 10 Rounds

	L	*C*	*R*
s	**0%**	0%	**53%**
t	0%	3%	**44%**

Last 10 Rounds

(d) Session 4

	L	*C*	*R*
s	**16%**	4%	**20%**
t	4%	7%	**49%**

First 10 Rounds

	L	*C*	*R*
s	**16%**	14%	**18%**
t	8%	12%	**32%**

Last 10 Rounds

(e) Session 5

Table 25: Communication Outcomes in Mediated-Direct Mechanism (Session Level): Joint Frequencies over Types and Actions in the First and Last 10 Rounds

E Language and Determinants of Behavior

We run separate regressions for each treatment to see whether the language we make available to subjects has a noticeable impact on their behavior. We find that for senders, regardless of the language, the principal drivers of the choice of sending message "t/R" are $t \times P$ under direct talk and t as well as $t \times P$ under mediated talk, confirming the findings from our pooled regression. Similarly, for receivers, we find that, like in the pooled regression, the principal driver of the choice of taking action L is S (i.e., receiving message "s/L"), and that the impact of that variable is more pronounced under mediated talk. This is independent of the language (except that in the direct declarative treatment the variable S is only marginally significant).

Table 26: Linear Probability and Probit Models: Senders in Direct-Talk Treatments

	Direct-Declaratives		Direct-Directives	
	(1)	(2)	(3)	(4)
Constant	0.142***	–	0.091***	–
	(0.016)	–	(0.015)	–
t	0.003	0.005	0.061*	0.061*
	(0.033)	(0.033)	(0.027)	(0.028)
P	0.116	−0.033	0.057**	0.025
	(0.064)	(0.044)	(0.021)	(0.021)
$t \times P$	0.320***	0.177**	0.355***	0.156***
	(0.100)	(0.065)	(0.030)	(0.013)
No. of Observations	2,537	2,537	2,891	2,891

Note: The dependent variable is an indicator for message “t/R”. Columns (1) and (3) report the coefficients from estimating the random-effects linear probability model for the corresponding treatments. Columns (2) and (4) report the *average marginal effects* from estimating the random-effects probit model for the corresponding treatments. Since the reported numbers are marginal effects, no constant term is included. Robust standard errors clustered at the session level are in parentheses. *** indicates significance at 0.1% level, ** significance at 1% level, and * significance at 5% level.

Table 27: Linear Probability and Probit Models: Senders in Mediated-Talk Treatments

	Mediated-Declaratives		Mediated-Directives		Mediated-Direct-Mechanism	
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.036***	–	0.042***	–	0.046***	–
	(0.008)	–	(0.007)	–	(0.014)	–
t	0.659***	0.426***	0.681***	0.434***	0.658***	0.405***
	(0.052)	(0.006)	(0.083)	(0.027)	(0.101)	(0.007)
P	−0.042*	−0.030	−0.071***	−0.035	−0.069**	−0.082**
	(0.021)	(0.041)	(0.021)	(0.034)	(0.023)	(0.319)
$t \times P$	0.270***	0.175***	0.254***	0.125***	0.288***	0.217***
	(0.040)	(0.030)	(0.057)	(0.026)	(0.087)	(0.060)
No. of Observations	2,537	2,537	2,832	2,832	2,655	2,655

Note: The dependent variable is an indicator for message “t/R”. Columns (1), (3), and (5) report the coefficients from estimating the random-effects linear probability model for the corresponding treatments. Columns (2), (4), and (6) report the *average marginal effects* from estimating the random-effects probit model for the corresponding treatments. Since the reported numbers are marginal effects, no constant term is included. Robust standard errors clustered at the session level are in parentheses. *** indicates significance at 0.1% level, ** significance at 1% level, and * significance at 5% level.

Table 28: Linear Probability and Probit Models: Receivers in Direct-Talk Treatments

	Direct-Declaratives		Direct-Directives	
	(1)	(2)	(3)	(4)
Constant	0.087***	–	0.024	–
	(0.025)	–	(0.016)	–
S	0.089*	0.087	0.208***	0.320***
	(0.045)	(0.054)	(0.038)	(0.073)
E	−0.009	−0.010	0.012	0.078
	(0.014)	(0.016)	(0.028)	(0.047)
$S \times E$	−0.009	−0.004	0.028	−0.044
	(0.037)	(0.025)	(0.024)	(0.040)
No. of Observations	2,537	2,537	2,891	2,891

Note: The dependent variable is an indicator for action L. Columns (1) and (3) report the coefficients from estimating the random-effects linear probability model for the corresponding treatments. Columns (2) and (4) report the *average marginal effects* from estimating the random-effects probit model for the corresponding treatments. Since the reported numbers are marginal effects, no constant term is included. Robust standard errors clustered at the session level are in parentheses. *** indicates significance at 0.1% level, ** significance at 1% level, and * significance at 5% level.

Table 29: Linear Probability and Probit Models: Receivers in Mediated-Talk Treatments

	Mediated-Declaratives		Mediated-Directives		Mediated-Direct-Mechanism	
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.053**	–	0.036***	–	0.017	–
	(0.018)	–	(0.007)	–	(0.013)	–
S	0.656***	0.392***	0.707***	0.351***	0.730***	0.363***
	(0.104)	(0.017)	(0.087)	(0.018)	(0.132)	(0.029)
E	−0.002	−0.011	−0.022*	−0.033	−0.007	−0.007
	(0.008)	(0.017)	(0.011)	(0.018)	(0.008)	(0.025)
$S \times E$	−0.009	0.003	0.080***	0.048**	0.017	0.007
	(0.034)	(0.024)	(0.011)	(0.018)	(0.030)	(0.034)
No. of Observations	2,537	2,537	2,832	2,832	2,655	2,655

Note: The dependent variable is an indicator for action L. Columns (1), (3), and (5) report the coefficients from estimating the random-effects linear probability model for the corresponding treatments. Columns (2), (4), and (6) report the *average marginal effects* from estimating the random-effects probit model for the corresponding treatments. Since the reported numbers are marginal effects, no constant term is included. Robust standard errors clustered at the session level are in parentheses. *** indicates significance at 0.1% level, ** significance at 1% level, and * significance at 5% level.

E.1 Level-k classification of individual subjects for each session

For each of the treatments, our level-k model makes a single prediction for levels 1 and above. As we have seen, there are substantial departures from that prediction. To get a clearer picture of the nature of these departures, here we classify individual subjects according to whether their behavior is best described as level-0, level-k with $k \geq 1$, or resists classification.

Table 30: Proportion of L_0, $L_{k\geq1}$, and Unclassified: Sender-Subjects

	Direct Declaratives			Direct Directives		
Session	L_0	$L_{k\geq1}$	Unclassified	L_0	$L_{k\geq1}$	Unclassified
1	0.20	0.60	0.20	0.00	1.00	0.00
2	0.00	0.80	0.20	0.10	0.80	0.10
3	0.14	0.57	0.29	0.11	0.89	0.10
4	0.00	1.00	0.00	0.00	0.80	0.20
5	0.25	0.75	0.00	0.00	1.00	0.00
Mean	0.12	0.74	0.14	0.04	0.90	0.06

Table 31: Proportion of L_0, $L_{k\geq1}$, and Unclassified: Receiver-Subjects

	Direct Declaratives			Direct Directives		
Session	L_0	$L_{k\geq1}$	Unclassified	L_0	$L_{k\geq1}$	Unclassified
1	0.00	0.50	0.50	0.00	0.50	0.50
2	0.00	0.60	0.40	0.00	0.60	0.40
3	0.00	0.71	0.29	0.00	0.33	0.67
4	0.00	0.50	0.50	0.00	0.30	0.70
5	0.00	0.38	0.62	0.00	0.50	0.50
Mean	0.00	0.54	0.46	0.00	0.45	0.55

In the case of direct talk, for both senders and receivers, strategies are different for L_0 and $L_{k\geq1}$. For each subject, we calculate the frequencies of observed choices (based on data from all rounds) that are consistent with a given level for that subject's role. Note that L_0 and $L_{k\geq1}$ strategies share some common components. This is the case when the sender's type is s or when the receiver observes message "s/L". Observed choices that are consistent with both L_0 and $L_{k\geq1}$ classifications, are counted toward the frequencies of both L_0 and $L_{k\geq1}$. For each subject, there will be one frequency for L_0 and another for $L_{k\geq1}$. We single out the more frequent one. If this higher frequency is no less than 70%, the subject is classified as belonging to that level. Otherwise, the subject is considered unclassified. For each session and each role, we calculate the proportions of L_0-subjects, $L_{k\geq1}$-subjects, and unclassified subjects. Tables 30 and 31 present the findings for, respectively, senders and receivers.

The classification is imperfect, with the degree of conformity with the level-k prediction varying both across sessions and across treatments. One characteristic of the classification that stands out is that fewer receiver subjects than sender subjects can be classified by our rule. On average, around

50% of receiver subjects cannot be classified as using a level-k strategy, whereas for senders more than 80% can be classified. In both cases, when subjects can be classified, they are overwhelmingly categorized as level $L_{k\geq 1}$ rather than level L_0 players, consistent with the notion that the level zero type is only a mental construct, the model used by the lowest level "real" type.

With mediated talk, for both senders and receivers, the strategies are the same at all levels. The classification is therefore dichotomous: subjects are either classified as $L_{k\geq 0}$ players or remain unclassified.

Table 32: Proportion of $L_{k\geq 0}$ and Unclassified: Sender-Subjects

	Mediated Declaratives		Mediated Directives		Mediated Direct Mechanism	
Session	$L_{k\geq 0}$	Unclassified	$L_{k\geq 0}$	Unclassified	$L_{k\geq 0}$	Unclassified
1	0.90	0.10	0.89	0.11	0.90	0.10
2	0.75	0.25	0.90	0.10	1.00	0.00
3	1.00	0.00	0.80	0.20	1.00	0.00
4	0.89	0.11	0.56	0.44	0.29	0.71
5	0.70	0.30	1.00	0.00	0.78	0.22
Mean	0.85	0.15	0.83	0.17	0.79	0.21

Table 33: Proportion of $L_{k\geq 0}$ and Unclassified: Receiver-Subjects

	Mediated Declaratives		Mediated Directives		Mediated Direct Mechanism	
Session	$L_{k\geq 0}$	Unclassified	$L_{k\geq 0}$	Unclassified	$L_{k\geq 0}$	Unclassified
1	0.60	0.40	0.67	0.33	0.80	0.20
2	0.50	0.50	0.60	0.40	1.00	0.00
3	0.67	0.33	0.50	0.50	0.67	0.33
4	0.78	0.22	0.33	0.67	0.29	0.71
5	0.40	0.60	1.00	0.00	0.44	0.56
Mean	0.59	0.41	0.62	0.38	0.64	0.36

We use the same 70% threshold. For each subject, we calculate the frequency of observed choices that are consistent with the $L_{k\geq 0}$ strategy of his/her role. If this frequency is no less than 70%, the subject is classified as a $L_{k\geq 0}$-type. Otherwise, he/she is considered unclassified. Tables 32 and 33 present the findings for, respectively, senders and receivers. As in the case of direct talk there is considerable variation across sessions and treatments, and senders are more frequently classified as $L_{k\geq 0}$ players than are receivers.

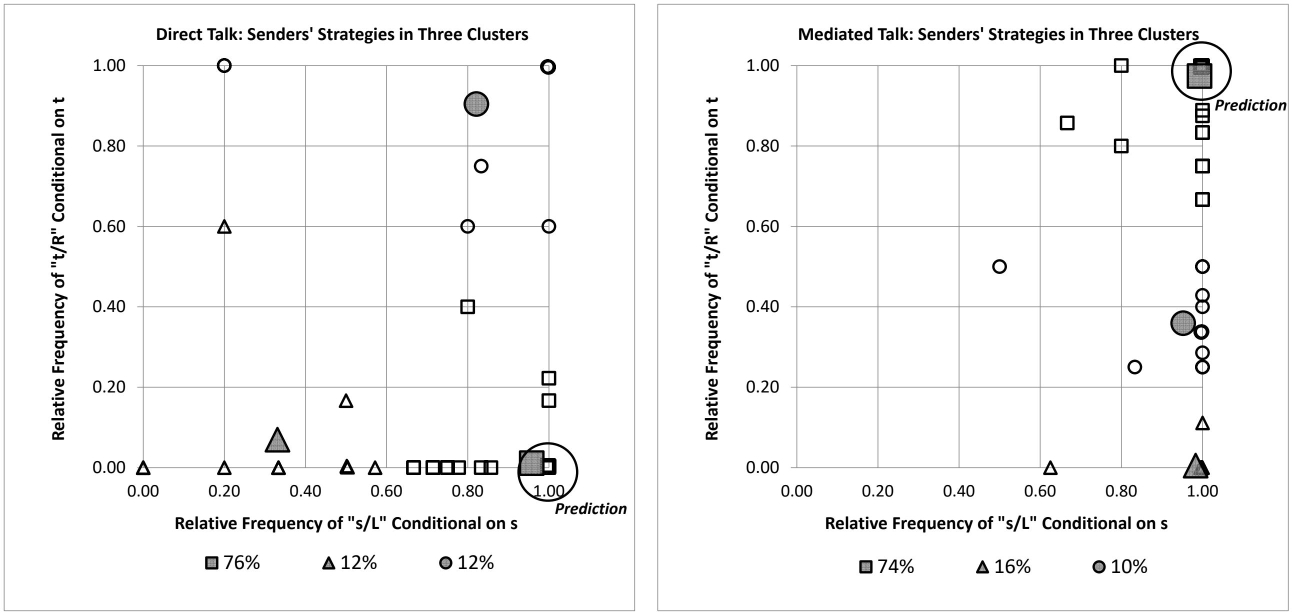
The finding that receiver strategies less frequently fit predicted level-k behavior is in line with our earlier observations about strategy choices: Under direct talk, receivers frequently use strategies that agree with the level-k prediction on the path of play, but not off the path of play. Under mediated talk, it is common for receivers to use strategies that can be rationalized in terms of treating mediated as direct talk – they either match the on-path prediction for direct talk or amount to credulous responses under direct talk.

E.2 Individual behavior: strategy choices and heterogeneity

In this section we take a closer look at the strategies that individuals adopt in the last 10 rounds. We ask which strategies feature prominently and how much heterogeneity there is in the strategies used.

To this end, we perform k-means clustering on subjects' strategies.[27] For each subject, we use their conditional choices in the last 10 rounds as proxies for their behavior strategies. The resulting observations, one for each subject, are partitioned into k clusters, with k pre-determined, based on the proximity of each observation to the center (mean) of a cluster. A larger value of k allows for a finer categorization of behavior, but may result in clusters with only a few observations. To balance these two considerations, we choose the value of k so that in each case the least frequent cluster contains approximately 10% of all observations.[28]

E.2.1 Sender strategies



(a) Direct Talk

(b) Mediated Talk

Figure 20: Three-Means Clustering: Senders

The 10% rule for the size of the minimal cluster results in having $k = 3$ for senders under both direct and mediated talk. In Figure 20 we plot the relative frequency of message "t/R" conditional on type t against the relative frequency of message "s/L" conditional on type s. Each marker in Figure 20(a) represents a subject in the two treatments of direct talk, and each marker in Figure

[27] MacQueen, James [1967], "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 281-297.

[28] This results in having each cluster contain at least 8.7% of all observations.

20(b) represents a subject in the three treatments of mediated talk. Cluster centers are shaded with gray. Markers with thickened borders represent multiple subjects whose choices coincide. In both figures the prediction of the theory is indicated by (the center of) a large circle.

There are 92 sender subjects under direct talk.[29] The □-cluster, with the highest proportion of observations (70 senders, 76%), closely matches the pooling on "s/L" strategy, ("s/L", "s/L"), where the first component indicates the message sent in state s and the the component the message sent in state t. The two remaining clusters each account for 12% of the observations. The ○-cluster corresponds to the separating strategy, ("s/L", "t/R"), and the △-cluster combines pooling on "s/L" with the flipped separating strategy ("t/R", "s/L").

In summary, under direct talk a large majority of senders conforms with the pooling-on-"s/L" prediction of the theory. The remaining senders either use the separating strategy, consistent with over-communication or combine pooling on "s/L" with the flipped separating strategy ("t/R", "s/L").

There are 136 sender subjects under mediated talk. The □-cluster, with the highest proportion of observations (100 senders, 74%), closely matches the separating strategy ("s/L", "t/R"). The △-cluster, with the second highest proportion of observations (16%), approximates pooling on "s/L". The ○-cluster, with the remaining observations (10%), can be viewed as a mixture of separation and pooling on "s/L". It is worth noting that none of subjects used either the pooling on "t/R" strategy ("t/R", "t/R") or the flipped separating strategy ("t/R", "s/L").

In summary, under mediated talk a large majority of senders conforms with the separating prediction of the theory, using the strategy ("s/L", "t/R"). The remaining senders either pool on message "s/L" using ("s/L", "s/L"), consistent with the theory prediction for direct talk, or end up somewhere between the strategies ("s/L", "t/R") and ("s/L", "s/L"). Thus, the modal strategy adopted by senders under mediated talk conforms with the theory prediction and departures from that prediction are consistent with some fraction of sender subjects treating mediated like direct talk.

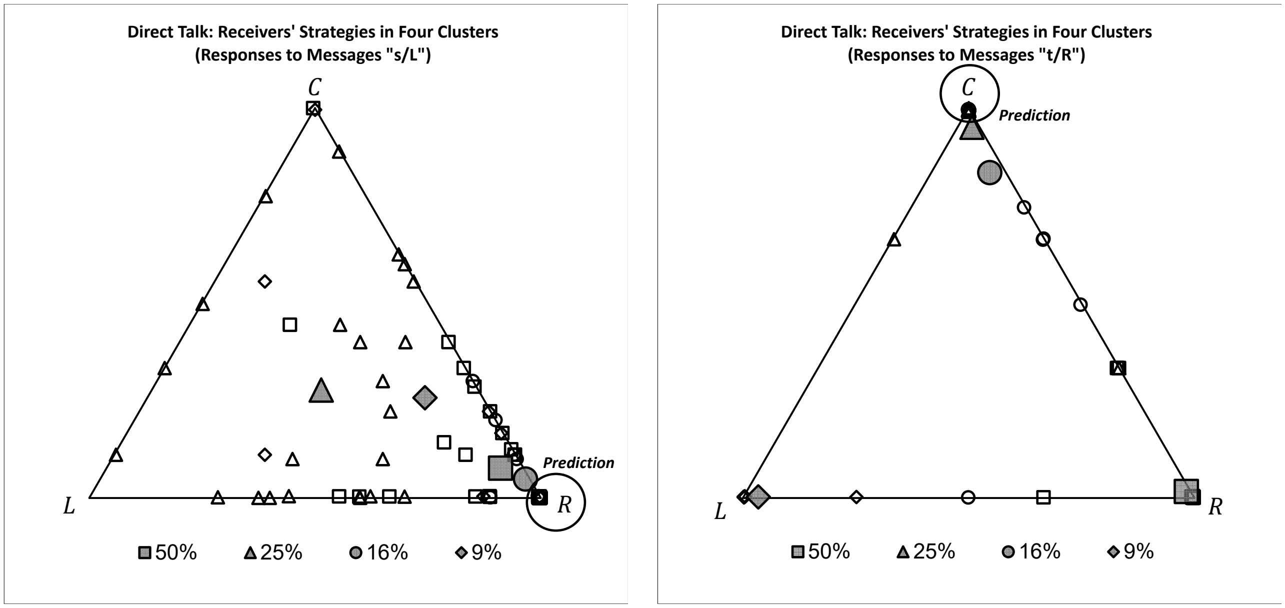
E.2.2 Receiver strategies

A behavior strategy of the receiver maps each message received to a distribution over the three actions, L, C, and R. We use two simplices to represent the empirical proxies of behavior strategies in the cluster analysis, one simplex for each message received. Figure 21 presents the analysis for direct talk. The 10% rule for the size of the least frequent cluster results in setting $k = 4$.

Note that each receiver subject corresponds to a *pair of markers* in Figure 21, since we cluster on strategies, not on choices. Cluster centers are shaded with gray. Thickened borders indicate multiple

[29] There is one sender subject for whom type t was not realized in any of the last 10 rounds. We use last-15-round data for that subject.

subjects adopting the same strategy. We represent (pure) receiver strategies by ordered pairs in which the first component describes the response to message "s/L" and the second component describes the response to message "t/R". The prediction of the theory is indicated by (the centers of) the large circles. It is worth recalling that the (R, C) prediction for direct talk is somewhat arbitrary in the second (C) component. Since senders are predicted to send only message "s/L", any receiver response to the "unsent" message "t/R" is part of a best reply. In particular, the receiver strategy (R, R) is both a best reply and supports the equilibrium prediction on the path of play.



(a) Relative Frequencies of Actions Conditional on "s/L"

(b) Relative Frequencies of Actions Conditional on "t/R"

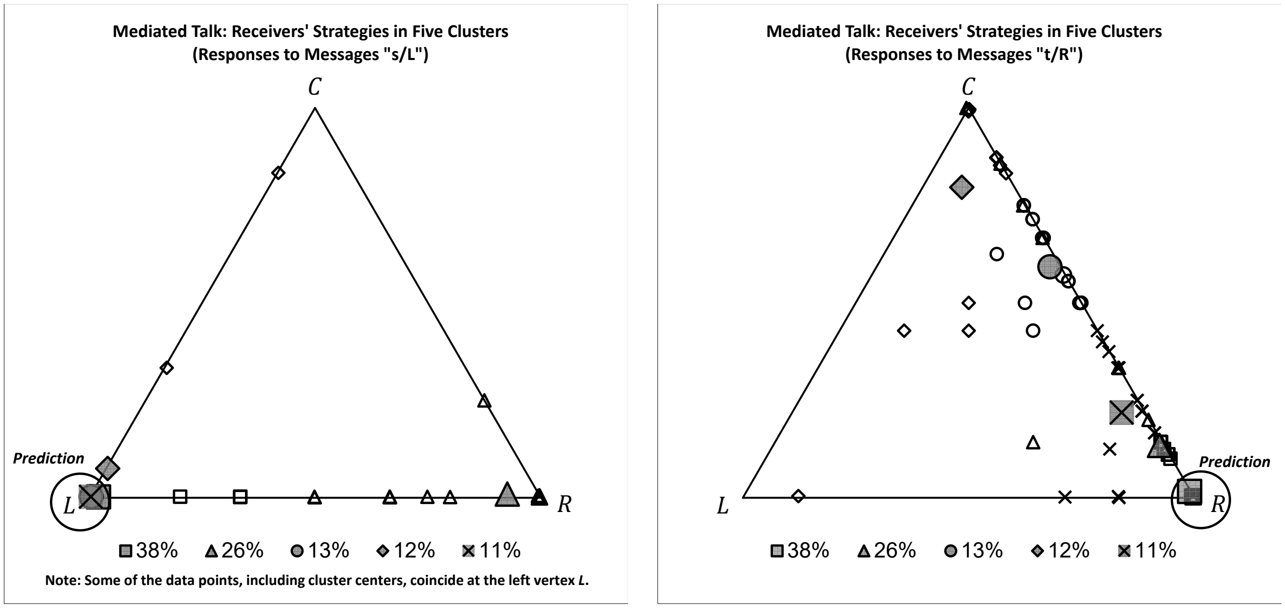
Figure 21: Four-Means Clustering: Receivers in Direct Talk

There are 92 receiver subjects under direct talk.[30] After the (on-path) message "s/L" two of the four clusters, the □-cluster and the ○-cluster, together accounting for 61 receivers (66% of the observations), very closely approximate the pooling response R predicted by theory. One of these two clusters, the □-cluster, accounting for 50% of the observations only matches the theory prediction after the on-path message "s/L". The other, ○-cluster (16% of observations), matches the theory prediction both on and off the path of play. The remaining two clusters give weight to all three actions after message "s/L" and respond to message "t/R" with with action C in the more frequent cluster (25% of observations) and with action L in the less frequent cluster (9% of observations).

In summary, a majority of receivers respond with action R to the message "s/L", consistent

[30] Theory predicts that message "t/R" is never sent under direct talk. In the data that message is observed infrequently and for some receiver subjects not at all in the last 10 rounds. For each of these 26 subjects, we go back five rounds at a time until we have data for the conditional action choices.

with the on-path prediction of the theory. This majority splits into a plurality who use the strategy (R, R), and a smaller group of subjects who use strategy (R, C). It is worth noting that there is no cluster corresponding to the separating strategy (L, C) for the receiver.



(a) Relative Frequencies of Actions Conditional on "s/L"

(b) Relative Frequencies of Actions Conditional on "t/R"

Figure 22: Five-Means Clustering: Receivers in Mediated Talk

Figure 22 presents the analysis for mediated talk. Our rule for the size of the least frequent cluster prescribes $k = 5$. Again, each subject corresponds to a *pair of markers*, cluster centers are shaded with gray, thickened borders indicate multiple subjects adopting the same strategy, and (the centers of) the large circles indicate the theory prediction.[31]

There are 136 receiver subjects under mediated talk. After message "s/L" four of the five clusters, accounting for 100 receivers (74% of the observations), very closely approximate the separating response L. One of these four clusters, the □-cluster, accounting for 38% of the observations closely matches separation after both messages "s/L" and "t/R". The most systematic deviation from separation is found in the △-cluster, which accounts for 26% of the observations. This cluster closely approximates the (R, R)-strategy, which is path equivalent to the theory prediction for direct talk. Both the ○-cluster (13%) and the ◇-cluster (12%), among pure strategies are nearest to (L, C), the best reply to the modal sender strategy for a receiver who perceives mediated talk as direct talk. Notice that of the nine strategies available to the receiver, six are not represented by any of the clusters. Among these, strategies that respond to message "s/L" with action C, that is (C, C), (C, L) and (C, R) are almost never used.

[31] Four receivers did not receive message "t/R" in any of the last 10 rounds. Using the last 15 rounds allows us to assign each of them to a cluster.

In summary, a large majority of receivers respond to message “s/L” with action L. A plurality use the strategy (L, R), consistent with the theoretical prediction. The other two strategies used with non-negligible frequency are (R, R) and (L, C). (R, R) is on-path equivalent to the theory prediction for direct talk. (L, C) is the best reply to the modal observed sender-strategy (recall that past sender behavior is part of the history information available to players) for a receiver who perceives mediated talk as direct talk. Thus, the bulk of strategies adopted by receivers under mediated talk conform with the theory prediction on the predicted path of play and departures from that prediction are consistent with some fraction of receiver subjects treating mediated like direct talk.

E.3 Initial behavior and behavior over time

In this section we take a look at initial behavior, examine how aggregate behavior evolves over time, and relate initial to terminal choices of strategies. We compare the data from mediated talk, aggregated over all sessions of the three mediated-talk treatment, with the data from direct talk, aggregated over all sessions of the two direct-talk treatments.

E.3.1 The evolution of sender behavior

Table 34 summarizes sender behavior in the mediated-talk and direct-talk treatments over the first 10 rounds, and compares observed with predicted behavior.

Table 34: Sender Behavior in Direct Talk vs. Mediated Talk

	“s/L”	“t/R”
s	**100%**	0%
t	**100%**	0%

Predicted

	“s/L”	“t/R”
s	**89%**	11%
t	**61%**	39%

First 10 Rounds

(a) Direct Talk

	“s/L”	“t/R”
s	**100%**	0%
t	0%	**100%**

Predicted

	“s/L”	“t/R”
s	**92%**	8%
t	21%	**79%**

First 10 Rounds

(b) Mediated Talk

In the first 10 rounds of the direct-talk treatments, type-t senders send message “s/L” 61% of the time. While this is significantly more often than message “t/R” ($p = 0.02$, Wilcoxon signed-rank test), the tendency for type-t senders to pool is not as pronounced in the initial rounds as in the terminal rounds. Initially, under direct talk there is some over-communication.

In the first 10 rounds of the mediated-talk treatments, type-t senders send message “t/R” 79% of the time, significantly more often than message “s/L” ($p < 0.001$, Wilcoxon signed-rank test). Type t-senders in the mediated-talk treatments separate by sending “t/R” significantly more often than do type t-senders in the direct-talk treatments ($p < 0.001$, Mann-Whitney test). Thus, under mediated talk initially there is already a strong tendency toward separation.

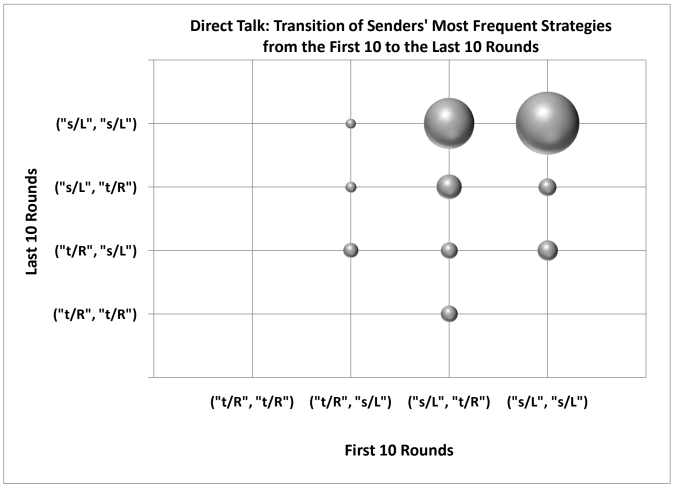


Figure 23: Sender strategies over time – direct talk

Figure 23 describes the strategies sender subjects adopt in direct talk and how the choices of strategies change from the initial 10 rounds to the terminal 10 rounds.[32] The strategies that dominate initially are (“s/L”, “s/L”)-pooling and (“s/L”, “t/R”)-separation in accordance with focal message use. Subjects who use the separating strategy initially, for the most part gravitate toward pooling on “s/L” in the end. None of the subjects start with pooling on “t/R”. The classification of terminal strategies closely mirrors the one identified by the k-means clustering analysis: There is a predominance of (“s/L”, “s/L”) pooling, followed by some focal and some flipped separation.

Figure 24 reports the evolution of sender strategies in mediated talk. There is a strong tendency toward separation that persists throughout. The only strategies used in the terminal rounds are focal separation, (“s/L”, “t/R”), and pooling on “s/L”. These are also by far the most common strategies used initially. There is some churning, with subjecst both moving in and out of using the focal separation strategy. None of the subjects use the pooling-on-“t/R” strategy or the flipped separating strategy (“t/R”, “s/L”) in the terminal rounds. The classification of terminal strategies mirrors the one identified by the k-means clustering analysis: There is a predominance of the focal separating strategy (“s/L”, “t/R”) and some pooling on “s/L”. Focal separation is predicted by the theory and pooling on “s/L” is consistent with some subjects treating mediated like direct talk.

[32] In this and and following bubble charts players are categorized by the strategy they use most frequently in the first (last) 10 rounds.

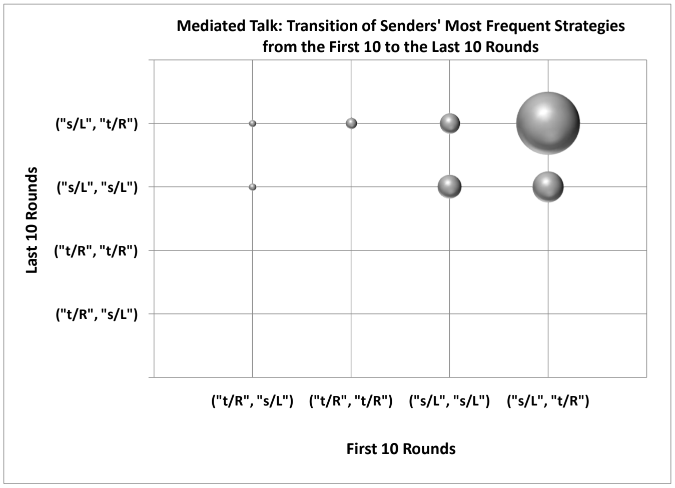


Figure 24: Sender strategies over time – mediated talk

E.3.2 The evolution of receiver behavior

Table 35 summarizes receiver behavior in the mediated-talk and direct-talk treatments over the first 10 rounds, and compares observed with predicted behavior.

	L	C	R
"s/L"	0%	0%	**100%**
"t/R"	0%	**100%**	0%

Predicted

	L	C	R
"s/L"	48%	6%	**46%**
"t/R"	2%	**55%**	43%

First 10 Rounds

(a) Direct Talk

	L	C	R
"s/L"	**100%**	0%	0%
"t/R"	0%	0%	**100%**

Predicted

	L	C	R
"s/L"	**83%**	2%	15%
"t/R"	3%	19%	**78%**

First 10 Rounds

(b) Mediated Talk

Table 35: Receiver Behavior in Direct Talk vs. Mediated Talk

Under direct talk, conditional on message "s/L" the frequency of R is 46% and conditional on message "t/R" it is 43% in the first 10 rounds. The frequency of action L after message "s/L" is 48%, and after message "t/R" the most frequent action is C, with a frequency of 55%. Thus initial receiver behavior under direct talk, departing from the theory prediction, is more consistent with separation than pooling.

In the first 10 rounds of the mediated-talk treatments 83% of receiver responses to message "s/L" are action L and 78% of receiver responses to message "t/R" are action R. Thus modal behavior in the initial periods of mediated talk conforms with the theory prediction. Foreshadowing behavior in the terminal ten rounds, two kinds of noteworthy of systematic departures from the theory under mediated talk stand out: the frequency of action R conditional on message "s/L" is 15% and the frequency of action C conditional on message "t/R" is 19%, while theory says that both should be 0%. As noted in our discussion of behavior in the last ten rounds, these departures are consistent with some subjects treating mediated talk as direct talk, either adopting equilibrium behavior for direct talk or best responding to the observed sender separation.

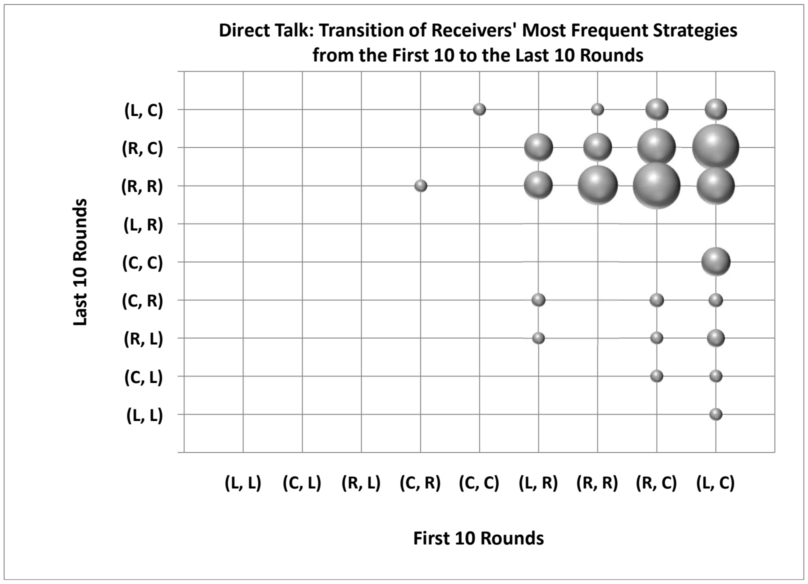


Figure 25: Receiver strategies over time – direct talk

Figure 25 describes the evolution of receiver strategies from the first to the last 10 rounds of direct talk. Initially, the four most prominent strategies are (L, C), (R, C), (R, R) and (L, R). The two strategies (R, C), (R, R) are consistent with the theory on the predicted path of play. Strategy (R, C) is fully consistent with the theory. The principal departure from the theory prediction is the use of the separating response (L, C). In the last 10 rounds this separating response largely disappears. The strategy (L, R) disappears entirely. The dominant strategies in the last 10 rounds are (R, R) and (R, C), the two strategies that agree with the theory prediction on the path of play. This classification of terminal strategies mirrors that from what we found with k-means clustering: a majority of receivers respond with action R following the on-path message "s/L".

Figure 26 describes the evolution of receiver strategies from the first to the last 10 rounds of mediated talk. Both initially and in the terminal 10 rounds the three most prominent strategies are (L, C), (R, R), and (L, R). There is some churning among these three strategies. Predominantly,

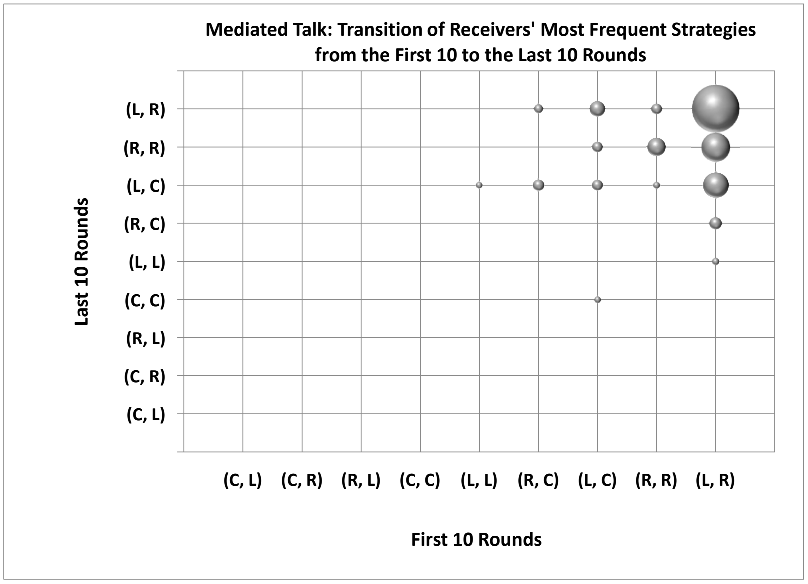


Figure 26: Receiver strategies over time – mediated talk

subjects use the separating response (L, R), both in the initial and the terminal 10 rounds, although there is some leakage from the initial to the terminal 10 rounds. (L, R) and (R, R), in this order, are also the most prominent terminal strategies according to our k-means clustering analysis. The separating strategy (L, R) matches the prediction of the theory; the strategy (R, R) is on-path equivalent to the theory prediction for direct talk; and, the strategy (L, C) amounts to credulous behavior for a receiver who perceives the game as direct talk. It is worth emphasizing that the strategy (L, C) is only rationalizable under direct talk. Hence, the receiver strategy gravitates toward the prediction of the theory, with departures that are explainable in terms of some subjects treating mediated talk like direct talk.

F Experimental Instructions - Mediated-Direct-Mechanism

Welcome to the experiment. This experiment studies decision making in pairs of individuals. The experiment will last approximately one and a half hours. There will be 60 rounds, and in each round you will be making one decision. Please read these instructions carefully. The cash payment you will receive at the end of the experiment will depend on the decisions you make.

Your Role and Your Pair

There are 20 participants in today's session. 10 of the participants will be randomly assigned to the role of a **Sender**. The other 10 participants will be randomly assigned to the role of a **Receiver**. Your role will remain fixed throughout the experiment. In each round a Sender will be paired with a Receiver so that 10 pairs will be formed. The pairing in each round is random. You will not learn the identity of the participant you are paired with, nor will that participant learn your identity, even after the end of the experiment.

Situations

There are two possible situations, **Situation *S*** and **Situation *T***. In each round, the computer randomly selects, with 50-50 chance, one of the two situations for a pair. The Sender will learn the situation. The Receiver will not learn the situation. Situations S and T differ by the rewards to the Sender and the Receiver, which will be explained below.

The Sender's Decision

If you are a Sender, after learning the situation, you decide on a **message** to send. You will be prompted to enter your choice of message. You have the choice to send message "S" by clicking the button marked "S" or to send message "T" by clicking the button marked "T." Once you click a button, your decision for that round has been made.

Action Recommendation

If the Sender sends message "S,"

- there is a 50% chance that action "L" is recommended to the Receiver; and
- there is a 50% chance that action "R" is recommended to the Receiver.

If the Sender sends message "T," then action "R" is always recommended to the Receiver.

The Receiver's Decision

If you are a Receiver, after seeing the action recommended to you, you decide which **action** to take. You will be prompted to choose one of the three actions "L," "C," or "R" by clicking one of the three buttons "L," "C," or "R." Once you click a button, all decisions for that round have been made, and your reward and the Sender's reward for that round is determined.

Your Reward

Figure 27(a) contains the decision screen for the Sender. Figure 27(b) contains the decision screen for the Receiver. The table in each screen contains the possible rewards. There are 6 cells in the reward table. The first number in a cell is the Sender's reward in HKD, and the second number is the Receiver's reward in HKD. **The selected situation and the Receiver's choice of action will determine which one of the six cells is used in the current round for rewards.** You will receive the reward number highlighted in blue in the relevant cell.

Information Feedback

At the end of each round, you will be provided with a summary of what happened in the round, including the selected situation, the Sender's choice of message, the action recommended to the Receiver, the Receiver's choice of action, and your reward for the round.

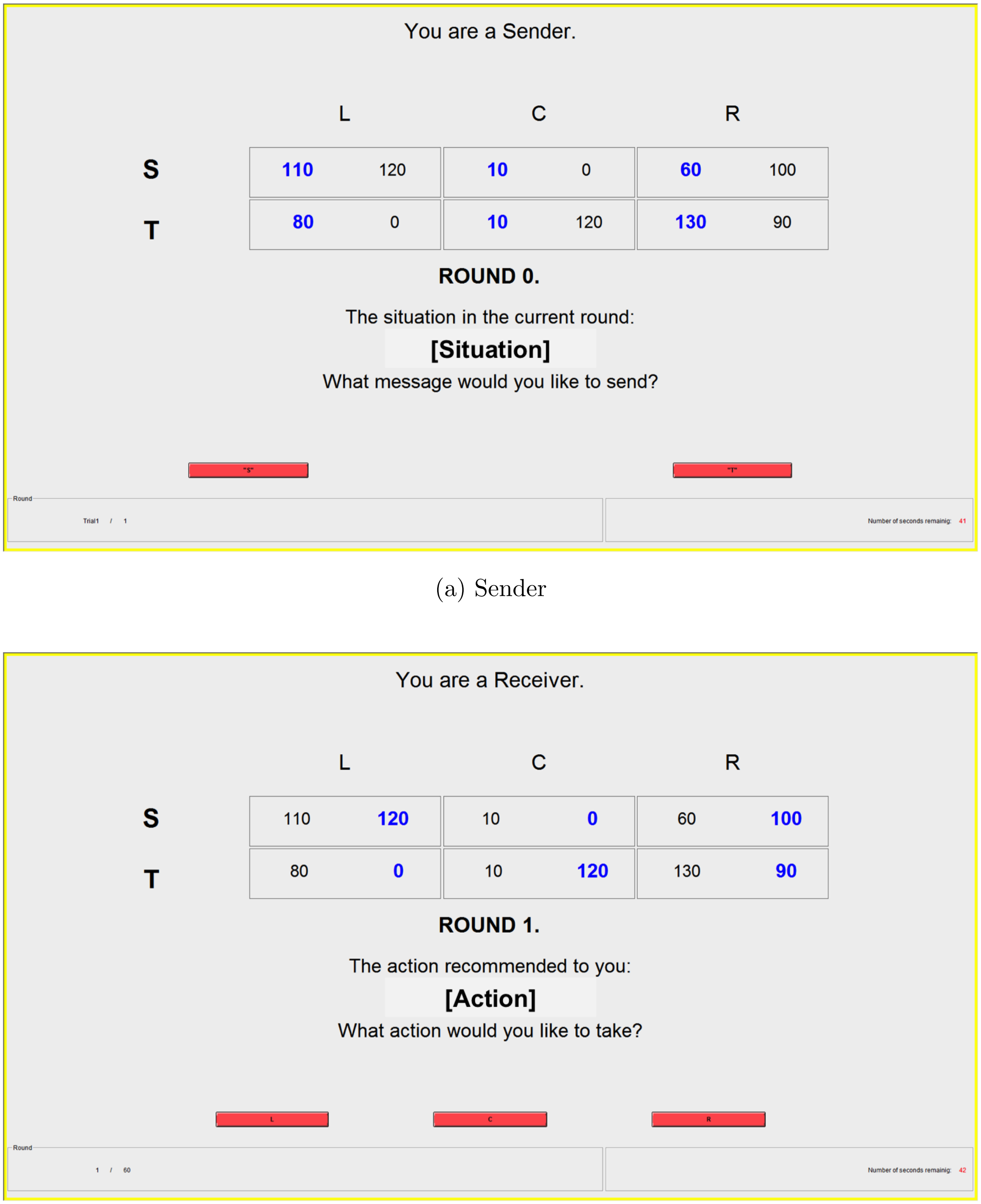
Your Cash Payment

The computer will randomly select 2 rounds out of the 60 to calculate your cash payment. (So it is in your best interest to take each round seriously.) Your total cash payment at the end of the experiment will be the average HKD you earned in the 2 selected rounds plus a 30 HKD show-up fee.

Quiz and Practice

To ensure your comprehension of the instructions, we will give you a quiz and a practice round. We will go through the quiz after you answer it on your own. You will then participate in 1 practice round. At the beginning of the practice round, you will be randomly assigned to the role of either a Sender or a Receiver. Your role in the official rounds is the same as that in the practice round. Once the practice round is over, the computer will tell you "The official rounds begin now!"

Administration



You are a Sender.

	L		C		R	
S	**110**	120	**10**	0	**60**	100
T	**80**	0	**10**	120	**130**	90

ROUND 0.

The situation in the current round:

[Situation]

What message would you like to send?

"S"
"T"
Round
Trial1 / 1
Number of seconds remainig: 41

(a) Sender

You are a Receiver.

	L		C		R	
S	110	**120**	10	**0**	60	**100**
T	80	**0**	10	**120**	130	**90**

ROUND 1.

The action recommended to you:

[Action]

What action would you like to take?

L
C
R
Round
1 / 60
Number of seconds remainig: 42

(b) Receiver

Figure 27: Decision Screens and Rewards

Your decisions as well as your monetary payment will be kept confidential. Remember that you have to make your decisions entirely on your own. Please do not discuss your decisions with any other participants.

Upon finishing the experiment, you will receive your cash payment. You will be asked to sign your name to acknowledge your receipt of the payment. You are then free to leave.

If you have any question, please raise your hand. We will answer your question individually. If there is no question, we will proceed to the quiz now.

Quiz

1. True or False: I will remain as a Sender or a Receiver in all 60 rounds of decision-making. Circle one: True / False
2. True or False: I will be paired with the same participant in the other role in all 60 rounds. Circle one: True / False
3. True or False: The Sender, but not the Receiver, will learn the situation selected by the computer. Circle one: True / False.
4. True or False: The Sender will be responsible for taking actions, while the Receiver will be sending messages. Circle one: True / False
5. True or False: If the Sender sends message "S," then action "L" is always recommended to the Receiver. Circle one: True / False
6. True or False: If the Sender sends message "T," then action "R" is always recommended to the Receiver. Circle one: True / False