Are We Strategically Naïve or Guided by Trust and Trustworthiness in Cheap-Talk Communication?

Xiaolin Li, Özalp Özer, Upender Subramanian
London School of Economics and Political Science, Department of Management, x.li166@lse.ac.uk
Naveen Jindal School of Management, The University of Texas at Dallas, {oozer, upender}@utdallas.edu

Abstract

Cheap-talk communication between parties with conflicting interests is common in many business and economic settings. Two distinct behavioral economics theories, the *trust-embedded model* and the *level-k model*, have emerged to explain how cheap talk works between human decision makers. The trust-embedded model considers that decision makers are motivated by non-pecuniary motives to be trusting and trustworthy. In contrast, the level-k model considers that decision makers are purely self-interested, but limited in their ability to think strategically. While both theories have been successful in explaining cheap-talk behaviors in separate contexts, they point to contrasting drivers for human behaviors. In this paper, we provide the first direct comparison of both theories within the same context. We show that, in a cheap-talk setting that well represents many practical situations, the two models make characteristically distinct and empirically distinguishable predictions. We leverage past experiment data from this setting to determine what aspects of cheap-talk behavior each model captures well, and which model (or combination of models) has better explanatory power and predictive performance. We find that the trust-embedded model emerges as the dominant explanation. Our results thus highlight the importance of investing in systems and processes to foster trusting and trustworthy relationships in order to facilitate more effective cheap-talk interactions.

Keywords: Behavioral Economics, Bounded Rationality, Cheap Talk, Level-k Thinking, Trust, Trustworthiness.*

History: Submitted January 2019. Revised Jun 2020, Oct 2020.

1 Introduction

In many business and economic settings, communication that is essentially *cheap talk* in nature may be used by a party with superior or proprietary information to influence decisions made by a less-informed party. For example, consumers receive product information from advertisers and

* The authors thank Ernan Haruvy, Tanjim Hossain, Sanjay Jain, Karen Zheng and participants in their presentation at the 2018 Behavioral Operations Management Conference, 2019 Behavioral IO and Marketing Symposium, 2019 Marketing Science Conference, 3rd Invitational Pricing Symposium, 2020 UTD BASS FORMS Conference and Johns Hopkins University for their constructive comments. The authors' names are listed in alphabetical order.

salespeople to make purchase decisions (e.g., Gardete 2013; Chakraborty and Harbaugh 2014), suppliers receive demand forecasts or soft orders from downstream firms to make capacity and production decisions (e.g., Terwiesch et al. 2005; Chu et al. 2016), individual investors receive investment recommendations from financial managers to make investment decisions (e.g., Michaely and Womack 1999; Angelova and Regner 2013), and patients receive drug and treatment procedure information from physicians to decide treatment plans (e.g., Schwartz et al. 2011; Guo et al. 2017). In these and other practical situations, the information communicated can be easy to distort and difficult to verify, and the communication does not create binding commitments. Consequently, "talk is cheap" and communication need not be truthful.

In fact, the interests of the two parties are never fully aligned and there is often a monetary incentive for one party to manipulate the other through cheap talk to obtain desired outcomes. For example, advertisers earn revenues and salespeople earn commissions from convincing consumers to buy; downstream firms realize higher sales and profit if they can ensure abundant upstream supply; financial managers make commissions if investors invest; and physicians bill for procedures performed and may receive perks from pharmaceutical companies. Nevertheless, cheap talk is widely used in such situations. At the same time, ample anecdotal evidence and empirical research have shown that cheap talk can be effective to coordinate decisions and improve outcomes in some instances, even as it has led to serious failures in other instances (e.g., Michaely and Womack 1999; Terwiesch et al. 2005; Özer et al. 2011; Schwartz et al. 2011; Brinkhoff et al. 2015). Therefore, it is important to understand how cheap talk works and when it can be effective.

Starting with the seminal work of Crawford and Sobel (1982), standard economic analysis has studied the effectiveness of cheap talk in different settings. Standard theory assumes that the *receiver* of the communication strategically anticipates how cheap-talk information may be distorted by the *sender* of the communication and responds to the information accordingly; similarly, the sender (she) strategically anticipates how the receiver (he) will be influenced by cheap talk and distorts the information accordingly. In equilibrium, these strategic behaviors interact and reinforce each other and determine whether and to what extent cheap talk can be effective. In this context, cheap talk is said to be *informative* if it (partially) reveals the sender's information, and *influential* if it affects the receiver's actions. Standard economic analysis provides bounds for how informative and influential cheap talk can be in a given setting.

However, experimental research has consistently found that there is systematic *over-communication*: cheap talk is more informative and influential than predicted by standard theory (e.g., Cai and Wang

2006; Mazar et al. 2008; Wang et al. 2010; Schwartz et al. 2011). In particular, informative and influential communication occurs even in cases where standard theory predicts that it should not (e.g., Forsythe et al. 1999; Özer et al. 2011; Spiliotopoulou et al. 2016). Another consistent finding is that there is substantial heterogeneity in behaviors, with senders differing in the extent to which they distort messages, and some receivers being more influenced than others (e.g., Cai and Wang 2006; Mazar et al. 2008; Wang et al. 2010; Özer et al. 2018).[1]

Motivated by these findings, two distinct behavioral economics theories have emerged to explain how cheap talk works between human decision makers. One theory is based on a positive view that individuals are guided not only by their self-interest, but also by non-pecuniary motives to be trusting and trustworthy in their interactions with others. Specifically, senders face a non-pecuniary disutility from lying due to factors such as adherence to social norms, guilt, maintaining positive self-image and other-regarding preferences (e.g., Gneezy 2005; Mazar et al. 2008; Lundquist et al. 2009; Erat and Gneezy 2012; Scheele et al. 2018). Consequently, they distort less than predicted by standard theory. Further, receivers place a higher belief in others acting in a trustworthy manner than predicted by standard theory for a "calculative" decision maker who uses Bayesian inference (e.g., Forsythe et al. 1999; Cain et al. 2010; Özer et al. 2011; Schwartz et al. 2011). Consequently, they are influenced more than predicted by standard theory. Moreover, an individual's disposition for being trusting or trustworthy is expected to be an intrinsic characteristic of that individual, gradually formed through life experiences, influenced by social norms, environment, and personal values (e.g., Moorman et al. 1993; Doney and Cannon 1997; Brehm and Rahn 1997; Hardin 2002; Özer et al. 2011; Beer et al. 2018; Özer and Zheng 2019). Consequently, there is heterogeneity in trusting and trustworthy behaviors (e.g., Ashraf et al. 2006; Gibson et al. 2013; Özer et al. 2018). Özer et al. (2011) formally propose and test a *trust-embedded* model of cheap talk that accounts for these factors. They show that it not only explains the observed effectiveness of cheap talk (while standard theory only predicts uninformative communication), but also fits and predicts individual participant-level (out-of-sample) behaviors well, and provides correct comparative statics predictions for how cheap-talk behavior is affected by changes in the experimental parameters.

The other behavioral theory is based on the somewhat grim view that individuals are primarily self-interested, but limited in their ability to think strategically in pursuit of their self-interest (e.g., Campbell and Kirmani 2000; Crawford 2003; Wang et al. 2010). Specifically, the *level-k model* of cheap talk considers that decision makers differ in their ability to think strategically. A

[1] We remark that these experiments are designed to minimize repeated interaction and reputation effects. Thus, the higher effectiveness of cheap talk cannot be explained by these mechanisms.

boundedly-strategic receiver does not fully anticipate the extent to which the sender, especially one who is more strategic than the receiver, may have distorted the message. Hence, such a receiver is more influenced than the fully-strategic receiver under standard theory. Similarly, a boundedly-strategic sender does not fully anticipate the extent to which the receiver may discount the message, and hence may distort less than the fully-strategic sender under standard theory. Consequently, boundedly-strategic thinking can lead to over-communication. Furthermore, heterogeneity in sender and receiver behaviors arises due to the differences in strategic ability. Cai and Wang (2006) find that the over-communication (relative to standard theory prediction) in their cheap-talk experiments could be explained by the level-k model. They find that the agent quantal response equilibrium (AQRE) model proposed by McKelvey and Palfrey (1998), in which players have correct beliefs about their opponents but deviate from their payoff maximizing decision due to bounded rationality, can also explain over-communication in their setting. Kawagoe and Takizawa (2009), however, find that only the level-k model can consistently explain over-communication in their cheap-talk experiments, whereas the AQRE model predicts only uninformative communication. Wang et al. (2010) use eye-tracking to capture how sender participants viewed payoff information in their cheap-talk experiment. Based on the level-k model, the authors structurally estimate each sender participant's level of thinking, and show that the payoffs that a participant pays most attention to is consistent with her estimated level of thinking.

Thus, both behavioral theories have been shown to explain cheap-talk behaviors better than standard theory, albeit in separate contexts. Both behavioral theories are also reasonable in many practical situations. For example, consumers might be influenced by salespeople because of their tendency to place trust based on their personal rapport with the salesperson (e.g., Doney and Cannon 1997; Schwartz et al. 2011); or, because their limited strategic thinking leads them to underestimate the salesperson's strategic intent in providing product information (e.g., Campbell and Kirmani 2000; Cain et al. 2010). Similarly, the salesperson may be relatively truthful either because of non-pecuniary cost of lying, or because of underestimating the consumer's ability to anticipate the strategic intent of the salesperson.

However, the two theories point to contrasting underlying drivers for behaviors and, therefore, lead to different implications for improving cheap-talk communication. For example, if behaviors are primarily driven by trust and trustworthiness, then measures to build trust such as reducing perceived vulnerabilities a receiver faces, or social uncertainties surrounding a sender can be effective in improving cheap-talk interactions (Özer and Zheng 2019). In contrast, if behaviors are primarily

driven by boundedly-strategic thinking, then what can be done to improve cheap-talk interactions can be quite limited. Instead, firms must reduce their reliance on cheap talk and / or make communication non-cheap-talk (through costly monitoring, verification, contracting or punishments).

Consequently, comparing the two theories within the same context is of theoretical and practical interest. Indeed, Joel Sobel, a pioneer of cheap-talk research, identified such a comparison between boundedly-strategic thinking and non-pecuniary motives as an important gap in the current understanding of cheap talk (Sobel 2013, page 409). Developing a more refined understanding of these behavioral drivers can help scholars and practitioners develop processes and establish more effective cheap-talk interactions that lead to profitable outcomes.

In this paper, we conduct the first direct comparison of both behavioral theories within the same unified context. Our objective is three-fold. First, to determine which behavioral model on its own has better explanatory power for observed behaviors in the same cheap-talk context. Second, to understand what aspects of behaviors each model capture well, and where each fares poorly. Third, to explore whether the performance of the individual models can be improved, including, by integrating them into a combined "hybrid" model. Overal our goal is to shed light on what aspects of behaviors each model captures well in the same context, and which model explains majority of the observed behaviors.

We leverage existing cheap-talk experiment data from Özer et al. (2018) to conduct the comparison; as we elaborate in §2, their context and experiments are particularly well-suited for comparing the two behavioral models and further well represent many practical situations, and the authors share this data with scholars (hence it is readily available). We show that while both models can potentially explain the over-communication phenomenon and heterogeneity in behaviors in this cheap-talk context, they lead to different implications for how people arrive at their decisions and, hence, predict distinct patterns of individual-level behaviors (i.e., even though, at first look, individuals' resulting decisions may look similar under both models, a careful analysis reveal how they are different). As a result, we are able to empirically distinguish between the model predictions based on how well they account for observed behaviors.

We structurally estimate both models, and compare their performance with respect to in-sample fit, out-of-sample predictions, and ability to recover the effect of an experimental manipulation.[2] We further examine several extensions to the models, including hybrid models. Overall, we find support

[2]Prior work that compare the explanatory powers of alternative behavioral theories in the same experiments using structural models of individual-level behaviors include Costa-Gomes and Crawford (2006) for guessing games (beauty contests), Crawford and Iriberri (2007) for auctions, Bostian et al. (2008) and Ho et al. (2010) for newsvendor models, and Kawagoe and Takizawa (2012) and Ho and Su (2013) for extensive-form games.

for the more positive view of human behavior as the dominant explanation; namely, that individuals are not only guided by their pecuniary self-interest, but also by their non-pecuniary motives of trust and trustworthiness. Thus, our results highlight the importance of designing processes to foster trusting and trustworthy relationships for cheap-talk interactions.

Our work adds to the growing behavioral literature in operations and marketing examining strategic interactions between parties in various business and economic environments. One literature stream has examined the role of boundedly-strategic thinking in settings different than ours. Researchers have used the cognitive hierarchy model to analyze competitive interactions between firms in the context of market entry (Goldfarb and Yang 2009; Goldfarb and Xiao 2011), pricing (Zhou et al. 2015) and inventory decisions (Chen et al. 2012; Feng and Zhang 2017; Cui and Zhang 2018), and also to study strategic interactions in presence of externalities in the context of reference-group effects amongst consumers of luxury goods (Amaldoss and Jain 2005, 2010) and participation in two-sided markets (Hossain and Morgan 2013). This literature stream has not examined cheap-talk communication between parties with conflicting interests.[3] Another literature stream has examined the role of trust and trustworthiness due to non-pecuniary motives in facilitating cheap-talk forecast sharing in a supply chain, between a single supplier and a retailer (Özer et al. 2011), between a single supplier and multiple retailers (Spiliotopoulou et al. 2016), and in a supply chain bridging cultures (Özer et al. 2014). Researchers have also examined the design of pecuniary contracts to incentivize truthful forecast sharing in a supply chain in the presence of non-pecuniary motives (Inderfurth et al. 2013; Spiliotopoulou et al. 2016). This literature stream on forecast sharing, however, has not examined the role of boundedly-strategic thinking. In contrast, the present paper examines whether non-pecuniary motives, more specifically trust and trustworthiness, or boundedly-strategic thinking, more specifically the level-k model, better explains cheap-talk behavior. Lastly, researchers have also examined how strategic interactions between collaborating parties with conflicting interests in other settings (without private information or cheap talk) is influenced by various non-pecuniary motives, such as fairness and reciprocity (Loch and Wu 2008; Katok and Pavlov 2013; Lim and Ham 2014; Cui and Mallucci 2016), reference dependence and loss aversion (Lim and Ho 2007; Ho and Zhang 2008; Katok and Wu 2009; Davis et al. 2014), and other types of social preferences (Kessler and Leider 2012; Lim and Chen 2014; Beer et al. 2018).

[3] We remark that Chen et al. (2012) and Cui and Zhang (2018) use the quantal-response equilibrium and cognitive hierarchy models, respectively, to explain the ordering decisions of multiple retailers served by a single supplier in a capacity allocation game. In this game, the retailers' ordering decisions constitute cheap talk. However, the focus of these studies is on the competitive interactions between retailers; the supplier's capacity is allocated (in proportion to their orders) according to a fixed allocation rule, and the strategic response of the supplier is not examined.

In what follows, in §2, we describe the cheap-talk game in Özer et al. (2018). In §3, we derive the behavioral predictions of the two behavioral economics models for this cheap-talk game and show that they lead to empirically distinguishable predictions. In §4, we discuss the experiment data and preliminary evidence supporting either model, and describe our model estimation and comparison approach. In §5, we present our results and observations. In §§6 and 7, we examine various model extensions. In §8, we conclude with a discussion on the implications of our findings.

2 A Context to Compare the Behavioral Models of Cheap Talk

To compare the behavioral models, we require a setting in which they make distinct predictions, i.e., each model predicts certain behaviors in that setting that can be uniquely explained only by that model. As noted in Sobel (2013), distinguishing between non-pecuniary preferences and boundedly-strategic behavior can be challenging as both can lead to some form of over-communication and heterogeneity in behaviors. We focus instead on the differences in the underlying decision processes and patterns of individual-level behaviors that lead to over-communication and heterogeneity.

Two features of the cheap-talk game in Özer et al. (2018) make it well-suited for our study. First, the sender's (pecuniary) payoff does not constrain the extent to which the sender will distort her message to mislead the receiver. This feature provides maximum scope to distinguish between a model that includes non-pecuniary factors that constrain behaviors, namely the trust-embedded model, and one that does not, namely the level-k model. Specifically, the sender's payoff is strictly increasing in the receiver's action. Consequently, the sender always prefers that the receiver takes the highest action regardless of the sender's private information. We say that the sender is *insatiable* since the sender is not "satisfied" no matter how high the receiver's action is. Another useful implication of this payoff structure is that the cognitive hierarchy model (an alternative model of boundedly-strategic thinking) predicts similar patterns of behavior as the level-k model (see §6.1). Hence, it is not necessary to separately consider such alternative models in this setting. Second, the sender and receiver strategy spaces are large: the number of possible sender messages is 81 and receiver actions is 181. Intuitively, a small strategy space limits the room for the predictions of the two models to differ and makes it difficult to reliably distinguish between their predictions, especially after allowing for random errors in participant decisions.[4] In contrast, a large strategy space provides more scope to observe differences in predicted behaviors.

Furthermore, the cheap-talk experiments in Özer et al. (2018) include an experimental manip-

[4] In a similar vein, Costa-Gomes and Crawford (2006) discuss the usefulness of larger strategy spaces in two-person guessing games to reliably differentiate between different models of non-equilibrium behavior.

ulation that provides an additional means to compare the models, namely, how well they predict and recover the effect of the manipulation. We formally derive the model predictions and elaborate rigorously on the aforementioned issues in §3 and throughout the paper. To our knowledge, only Özer et al. (2018) use cheap-talk experiments that incorporate an insatiable sender and large strategy spaces (see Appendix B for a comparison of the experimental setups in prior cheap-talk experiments). Therefore, their setting and data provides a natural and useful starting point for our study. We describe their cheap-talk game below.

Cheap-talk game in Özer et al. (2018). A supplier (e.g., P&G) shares information about its product's market (e.g., sales) potential with a retailer (e.g., Kroger) to influence how much store resources (e.g., shelf space) the retailer allocates to the supplier's product. The supplier is better informed about the market potential than the retailer. However, the information it shares is cheap talk. Moreover, the supplier prefers that its prodict is allocated the maximum amount of available store resources, irrespective of the product's actual market potential, because higher store resources result in higher sales and profits for the supplier. In contrast, the retailer prefers to allocate store resources depending on the market potential as store resources are costly. Knowing this fact, the supplier has an incentive to manipulate the market information it communicates to the retailer.

Formally, the market potential for the supplier's product is uncertain, given by $q = \xi + \epsilon$, where ξ is a positive integer uniformly distributed over $\left[\underline{\xi}, \bar{\xi}\right]$, and ϵ is an integer that is uniformly distributed over $[-e, e]$. The variable ξ is known to the supplier but not to the retailer, and represents the supplier's private demand information; the other variable ϵ is not known to either party, and represents the inherent market uncertainty. The sequence of events is as follows. The supplier observes its private demand information ξ and sends a report $\hat{\xi} \in \left[\underline{\xi}, \bar{\xi}\right]$ to the retailer. Then, the retailer decides the service level $a \in [0, \bar{a}]$, which represents the level of store resources allocated to promote the supplier's product. Then, market uncertainty ϵ and hence demand is realized. Demand for the product is given by $D(a) = qa$; thus, the effectiveness of retail service a in stimulating demand depends on the product's market potential q. The supplier delivers the product to the retailer at a profit margin s, and the retailer sells the product to end consumers at a retail margin r. The retailer's cost of service is $C(a) = \frac{1}{2}ca^2$. The supplier's and retailer's profits are respectively:[5]

$$\Pi_S = sqa = s(\xi + \epsilon)a, \tag{1}$$

[5]We adapt the model notation in Özer et al. (2018) to suit our context. In their notation, the supplier's payoff is wqs and the retailer's payoff is $(r - w)qs - \frac{1}{2}ks^2$, where s is the retailer's service level, r is the retail price, w is the wholesale price (and margin), and k is the cost parameter.

$$\Pi_R = rqa - \frac{1}{2}ca^2 = r\left(\xi + \epsilon\right)a - \frac{1}{2}ca^2, \tag{2}$$

Note that the retailer can benefit from knowing the supplier's private information, since service is costly and its effectiveness depends on the product's market potential. However, the supplier benefits from inducing as high a service decision as possible regardless of the actual market potential, since its profit is strictly increasing in the retailer's service level a. Thus, the interests of the two parties are not aligned. Moreover, communication regarding the product's market potential is essentially cheap talk (i.e., the supplier's report of ξ is costless, non-binding and not verifiable).

Since our interest extends beyond the above distribution channel setting, in the rest of the paper, we abstract away from this specific setting; specifically, we adopt the terminology of a sender-receiver cheap-talk game, denoting the supplier as the sender, and the retailer as the receiver. We also discuss below other cheap-talk settings with similar payoff structure, i.e., cheap talk by an insatiable sender. Thus, our findings can be of interest beyond the above distribution channel setting.

Other settings with cheap talk by an insatiable sender. We remark that cheap talk by an insatiable sender can represent many other practical situations. For example, online platforms (e.g., Airbnb, Amazon, eBay) are better informed about market conditions than sellers or service providers on their platform, and advise them on improving product or service quality to increase demand. A salesperson is better informed about his or her sales territory than management, and requests costly marketing support such as advertising or promotions to aid the sales process. Car mechanics are better informed about the condition of their customers' cars, and inform customers about how extensive repairs need to be. A physician is better informed about a patient's condition, and helps the patient decide between different treatment plans. In these examples, the sender (supplier, platform, salesperson, mechanic, physician) always benefits from inducing the receiver (retailer, seller, management, car owner, patient) to make a substantially "large" decision (in service level, service quality, marketing support, repair cost, intensity of or degree of medical treatment) regardless of the true need for the receiver to do so.

3 Predictions from Standard and Behavioral Economic Models

We first present the predictions of the standard theory model for the (aforementioned) cheap-talk game in Özer et al. (2018). These predictions serve as the benchmark for determining what constitutes over-communication in this context. We then develop and compare the theoretical predictions from the two behavioral economic models. We show that both models predict over-communication relative to standard theory predictions, and heterogeneity in the extent to which senders distort

their message and receivers are influenced by sender messages. However, they also predict different patterns of individual-level behaviors leading to over-communication and heterogeneity, due to the differences in the underlying decision-making processes.

Before, we proceed, it is useful to define the receiver's optimal strategies under two extreme scenarios, namely, when the receiver believes $\hat{\xi}$ is the true information and when he believes $\hat{\xi}$ is uninformative. Let $a_I\left(\hat{\xi}\right)$ denote the receiver's *fully-believing* optimal strategy when he believes $\hat{\xi}$ is truthful. Let a_{NI} denote the receiver's *fully-disbelieving* optimal strategy when he believes $\hat{\xi}$ is uninformative. We have

$$a_I\left(\hat{\xi}\right) = \arg\max_a \; r\mathbf{E}\left[q \mid \hat{\xi} = \xi\right] a - \frac{1}{2}ca^2 = \frac{r}{c}\hat{\xi}, \tag{3}$$

$$a_{NI} = \arg\max_a \; r\mathbf{E}\left[q\right] a - \frac{1}{2}ca^2 = \frac{r}{c}\frac{\left(\bar{\xi}+\underline{\xi}\right)}{2}. \tag{4}$$

3.1 Standard Theory Predictions

We restate the standard theory predictions from Özer et al. (2018). Let $\hat{\xi}\left(\xi\right)$ denote the sender's strategy after observing ξ. Let $a\left(\hat{\xi}\right)$ denote the receiver's strategy upon receiving the sender's message $\hat{\xi}$. Özer et al. (2018) show that all perfect Bayesian equilibria of this cheap-talk game are uninformative and uninfluential.

Theorem 1. *(Adapted from Özer et al. 2018) In any perfect Bayesian equilibrium, the sender's report $\hat{\xi}\left(\xi\right)$ is uncorrelated with ξ, and the receiver's action is $a\left(\hat{\xi}\right) = a_{NI}$, independent of $\hat{\xi}$.*

Theorem 1 is a consequence of the sender being insatiable (i.e., sender's payoff being strictly increasing in receiver's action). Intuitively, regardless of her actual information ξ, an insatiable sender always prefers to send a message that will induce the highest action that the receiver is willing to take. Consequently, the sender's message is uninformative. In equilibrium, the receiver correctly anticipates the sender's behavior and always takes the fully-disbelieving action a_{NI}, ignoring the sender's messages. Thus, standard theory predicts uninformative and uninfluential communication.

3.2 The Trust-embedded Model and Predictions

The trust-embedded model, originally introduced by Özer et al. (2011), incorporates individuals' non-pecuniary preferences to be trusting and trustworthy. In the context of communication games, trustworthiness is defined as the sender's tendency not to manipulate her message to her own monetary benefit at the expense of the receiver, and trust is defined as the receiver's tendency to rely on the sender's message to make decisions, even though the message may have been manipulated, exposing the receiver to significant monetary loss (see Özer and Zheng 2019 for a comprehensive

discussion of these definitions). We first describe the components of this model, and then obtain the predictions for our specific context.

The trust-embedded model differs from the standard economic model in three respects. First, a sender incurs a disutility from lying. Researchers have shown that some people are averse to lying in economic interactions even at the expense of their monetary gain (e.g., Gneezy 2005; Mazar et al. 2008; Lundquist et al. 2009; Erat and Gneezy 2012; Scheele et al. 2018), in effect experiencing an intrinsic psychological cost from lying. The lying cost can be attributed to factors such as adherence to social norms, maintaining positive self-image, avoiding hurting others, and other social motives. Furthermore, the lying cost is increasing in the "size of the lie" (e.g., Mazar et al. 2008; Gneezy et al. 2018; Scheele et al. 2018).

Accordingly, the trust-embedded model assumes that a sender incurs a lying cost $G\left(\left|\hat{\xi}-\xi\right|;\gamma\right)$, where $\left|\hat{\xi}-\xi\right|$ is the magnitude of the lie and $\gamma \geq 0$ denotes the sender's *lying cost* type. The lying cost function $G(\cdot)$ is strictly convex and increasing in the magnitude of the lie. Let $g\left(\left|\hat{\xi}-\xi\right|;\gamma\right) = \frac{\partial G(|\hat{\xi}-\xi|;\gamma)}{\partial \hat{\xi}}$ denote the marginal cost of lying. The lying cost type γ indexes the lying cost such that: (i) a sender incurs no lying cost if $\gamma = 0$, i.e., $G\left(\left|\hat{\xi}-\xi\right|;\gamma=0\right)=0$; and (ii) the sender's marginal cost of lying $g\left(\left|\hat{\xi}-\xi\right|;\gamma\right)$ is strictly increasing in γ for all $\left|\hat{\xi}-\xi\right|>0$. We impose the regularity condition that $G\left(\left|\hat{\xi}-\xi\right|;\gamma\right)$ is twice continuously differentiable for $\xi,\hat{\xi}\in\left[\underline{\xi},\bar{\xi}\right]$; hence, in particular, $g(0;\gamma)=0$. Let $g^{-1}(\cdot;\gamma)$ denote the inverse of the marginal lying cost, i.e., if $g(x;\gamma)=y$ then $g^{-1}(y;\gamma)=x$, where $x,y\geq 0$. In our empirical application (described in the following section), we use a quadratic lying cost $G(x;\gamma)=\frac{1}{2}\gamma x^2$; hence $g(x;\gamma)=\gamma x$ and $g^{-1}(y;\gamma)=\frac{y}{\gamma}$.

Second, the trust-embedded model assumes that a receiver follows a relatively simple (non-Bayesian) belief-updating rule that reflects his intrinsic tendency to trust. The standard economic model requires that a receiver follows Bayes rule to update his belief about ξ given $\hat{\xi}$, anticipating the sender's communication strategy for each ξ. However, such a belief-updating rule is complex even in simple communication games. Instead, human decision makers have been known to adopt simpler non-Bayesian belief-updating rules (e.g., Kahneman and Tversky 1982). Moreover, decision makers have been found to place more trust in others than predicted for a rational "calculative" decision maker, a behavior attributed to their inherent tendencies to trust (e.g., Cain et al. 2010; Özer et al. 2011; Schwartz et al. 2011). For example, receivers in communication games have been found to be overly trusting, assigning higher subjective probabilities of senders' messages being truthful than should be expected from a rational Bayesian decision maker (e.g., Sheremeta and Shields 2013; Jin et al. 2018) and being influenced by messages even in games where standard equilibrium predicts

that they should disregard all messages (e.g., Forsythe et al. 1999; Sánchez-Pagés and Vorsatz 2007; Özer et al. 2011; Spiliotopoulou et al. 2016) including in Özer et al. (2018).

Accordingly, the trust-embedded model assumes that a receiver follows a trust-based belief-updating rule that can be expressed as follows: given the message $\hat{\xi}$, a receiver believes that ξ has the same distribution as $\alpha_R\hat{\xi} + (1-\alpha_R)\,\xi^T$, where $\alpha_R \in [0,1]$ denotes the receiver's *trust type* and ξ^T follows the distribution of ξ truncated on $\left[\underline{\xi}, \hat{\xi}\right]$, namely, a uniform distribution. Essentially, the receiver believes with probability α_R that $\hat{\xi}$ is truthful, and with probability $1-\alpha_R$ that $\hat{\xi}$ is inflated but otherwise uninformative such that ξ may be any number in $\left[\underline{\xi}, \hat{\xi}\right]$ with equal probability. Here, α_R is interpreted as the receiver's tendency to trust. If α_R =1, then the receiver fully trusts that the sender is truthful and believes $\xi = \hat{\xi}$. If $\alpha_R = 0$, then the receiver believes the message is always inflated and $\xi \sim Unif\left[\underline{\xi}, \hat{\xi}\right]$.

Lastly, the trust-embedded model assumes that the sender has a belief about the receiver's trust type α_R, denoted by α_S with cdf $H(\cdot)$, that reflects the sender's belief about being trusted; for instance, $H(\cdot)$ may be based on her past experiences of being trusted.

Thus, under the trust embedded model, the expected payoff for a receiver of type α_R is

$$\Pi_R\left(a, \hat{\xi}; \alpha_R\right) = r\mathbf{E}\left[\xi \mid \hat{\xi}, \alpha_R\right] a - \frac{1}{2}ca^2, \tag{5}$$

where the receiver follows the trust-based belief-updating rule:

$$\begin{aligned}\mathbf{E}\left[\xi \mid \hat{\xi}, \alpha_R\right] &= \alpha_R\hat{\xi} + (1-\alpha_R)\mathbf{E}\left[\xi \mid \xi \le \hat{\xi}\right], \\ &= \frac{(1+\alpha_R)\,\hat{\xi} + (1-\alpha_R)\,\underline{\xi}}{2}.\end{aligned} \tag{6}$$

Note that if $\alpha_R = 1$, then $\mathbf{E}\left[\xi \mid \hat{\xi}, \alpha_R\right] = \hat{\xi}$ since the receiver believes $\xi = \hat{\xi}$. If $\alpha_R = 0$, then $\mathbf{E}\left[\xi \mid \hat{\xi}, \alpha_R\right] = \frac{1}{2}\left(\hat{\xi} + \underline{\xi}\right)$ since the receiver believes $\xi \sim Uniform\left[\underline{\xi}, \hat{\xi}\right]$. Let $a^*\left(\hat{\xi}; \alpha_R\right)$ denote the receiver's optimal strategy.

The expected payoff for a sender with lying cost type γ and belief $H(\alpha_S)$ about the receiver's trust type is

$$\Pi_S\left(\hat{\xi}, \xi; \gamma, H(\alpha_S)\right) = s\xi\mathbf{E}\left[a^*\left(\hat{\xi}; \alpha_R\right) \mid H(\alpha_S)\right] - G\left(\left|\hat{\xi} - \xi\right|; \gamma\right), \tag{7}$$

Let $\hat{\xi}^*(\xi; \gamma, H(\alpha_S))$ denote the sender's optimal strategy.

We derive the optimal sender and receiver strategies from Equations (5) and (7). From Equation (6), we observe that the receiver's belief is increasing in the sender's message. Thus, in general, the sender's message influences the receiver, and the receiver's action is increasing in the sender's message. Accordingly, the sender can have an incentive to inflate the message depending on her

lying cost. The following theorem describes the sender's and receiver's optimal strategies. All proofs are deferred to Appendix A.

Theorem 2. *Under the trust-embedded model, the sender's optimal strategy is*

$$\hat{\xi}^*\left(\xi;\gamma,H\left(\alpha_S\right)\right)=\min\left\{\bar{\xi},A\left(\gamma,\bar{\alpha}_S\right)\cdot\xi\right\},$$

where $\bar{\alpha}_S=\int\alpha_S dH\left(\alpha_S\right)$ *and* $A\left(\gamma,\bar{\alpha}_S\right)=1+g^{-1}\left(s\frac{r}{c}\left(\frac{1+\bar{\alpha}_S}{2}\right)\xi;\gamma\right)$. *The receiver's optimal strategy is*

$$a^*\left(\hat{\xi};\alpha_R\right)=\frac{r}{c}\left[\frac{\left(1+\alpha_R\right)\hat{\xi}+\left(1-\alpha_R\right)\underline{\xi}}{2}\right].$$

Theorem 2 finds that the sender inflates the message by a factor $A\left(\gamma,\bar{\alpha}_S\right)$ or up to $\bar{\xi}$, whichever is lower; where $\bar{\alpha}_S$ denotes the sender's *average belief* about the receiver's trust. We refer to $A\left(\gamma,\bar{\alpha}_S\right)$ as the sender's *trustworthiness factor*; noting that lower the trustworthiness factor, lower the extent to which the sender inflates the messages, and hence more trustworthy the sender. The trustworthiness factor, and therefore the extent to which a sender distorts the message, can vary across individuals. We observe that $A\left(\gamma,\bar{\alpha}_S\right)$ is decreasing in the sender's lying cost type γ since the marginal cost of lying $g\left(x;\gamma\right)$ is increasing in γ and in x; $A\left(\gamma,\bar{\alpha}_S\right)$ is also increasing in her average belief $\bar{\alpha}_S$ about the receiver's trust. For receivers, the extent to which a receiver's action is influenced by the sender's message is increasing in his trust type α_R. If $\alpha_R=1$, then the receiver fully believes the sender's message to be true ($\xi=\hat{\xi}$) and his optimal action is the fully-believing action $a_I\left(\hat{\xi}\right)$. If $\alpha_R=0$, then the receiver's optimal strategy is $a\left(\hat{\xi};\alpha_R=0\right)=\frac{r}{2c}\left(\hat{\xi}+\underline{\xi}\right)$; the receiver is still influenced by the sender's message because he believes that the true information is less than the received message. For $\alpha_R\in(0,1)$, the extent to which the receiver is influenced is between these two extremes.

Thus, the trust-embedded model predicts distorted, yet informative and influential communication. Hence, there is over-communication relative to standard theory predictions. Moreover, the trust-embedded model predicts heterogeneity in the extent to which senders distort messages and receivers are influenced by messages depending on their intrinsic trustworthiness and trust.

3.3 Level-k Model and Predictions

The level-k model of boundedly-strategic behavior has been used to explain deviations from standard theory in a wide variety of games (see Crawford et al. 2013 and Georganas et al. 2015 for recent surveys). It was originally proposed for normal-form complete information games by Nagel (1995) and Stahl and Wilson (1994, 1995), and subsequently adapted to a cheap-talk setting (a dynamic game of incomplete information) by Crawford (2003) and Cai and Wang (2006). The level-k model

departs from the standard economic model by assuming that players are limited in their ability to anticipate the strategic behaviors of others. Specifically, players are distinguished by their *level of thinking*: a type Lk player exhibits a level of thinking $k \geq 0$. The $L0$ player is non-strategic, representing naive or unsophisticated play. Each higher level player is more sophisticated, yet only boundedly so; anticipating the strategic behaviors of lower levels, but not of levels higher than oneself.

The level-k model starts with the specification of $L0$ player behavior. In complete information games, $L0$ players are specified as uniformly randomizing across all actions, i.e., the $L0$ player is unable to evaluate any consequence of her actions, strategic or otherwise, and chooses an action arbitrarily. In adapting the level-k model to cheap-talk games, researchers have specified the $L0$ sender as naively revealing her actual information (Crawford 2003; Cai and Wang 2006; Kawagoe and Takizawa 2009; Ellingsen and Östling 2010; Wang et al. 2010), i.e., when asked to communicate her information, the sender fails to consider how the receiver will respond to her message and how the receiver's action will affect her payoff, and simply reports her actual information. Further, an $L0$ receiver is specified as credulously believing the sender's message to be true. As Crawford et al. (2013) (pg. 51) explain: "it would be behaviorally odd if a [receiver's] strategically naive assessment of a message did not initially favor its literal interpretation". In §4.2, we provide preliminary evidence from the experiment data supporting this assumption. It is important to note that while the $L0$ behaviors resemble trusting and trustworthy behaviors, the underlying driver for these behaviors under the level-k model is the inability to think strategically. Further, as explained later in §6.2, in the context of a cheap-talk game, specifying $L0$ players as naively randomizing their actions does not lead to useful predictions. In §6.2, we also examine a level-k model in which a *fraction* of $L0$ players may be naively randomizing.

The behavior of higher level players is obtained through iterative thinking about the strategies of lower level players. Each higher type believes that other players are of the immediately lower type. Therefore, $L1$ best responds to $L0$ behavior, $L2$ best responds to $L1$ behavior and so on. Later, in §6.1, we discuss the cognitive hierarchy model proposed by Camerer et al. (2004), in which a higher-level type believes that the other player is drawn from a distribution of lower types.

Note that $L0$ receivers (who are believing) are in effect best responding to $L0$ senders (who are truth-revealing). As a result, it is not necessary to separately consider that $L1$ receivers best respond to $L0$ senders as doing so does not give rise to newer predicted behaviors. For example, an $L1$ receiver best responding to an $L0$ sender would behave identical to an $L0$ receiver; moreover,

an $L1$ sender and an $L2$ sender (best responding to $L0$ and $L1$ receivers, respectively) would have identical behaviors, and so on. To avoid this duplication of behaviors, we follow Cai and Wang (2006) and Wang et al. (2010) in reorganizing the levels of thinking for receivers as follows: an Lk receiver believes the sender is Lk (not $L(k-1)$); an Lk sender still believes the receiver is $L(k-1)$.

For an Lk sender, let $\Pi_{Sk}\left(\hat{\xi},\xi\right)$ denote the expected payoff given the true state ξ and her message $\hat{\xi}$; let $\hat{\xi}_k(\xi)$ denote her best response message given ξ; and, let $\Xi_k = \left\{x : \hat{\xi}_k(\xi) = x \text{ for some } \xi \in \left[\underline{\xi},\bar{\xi}\right]\right\}$ denote the (feasible) set of messages she might send for some $\xi \in \left[\underline{\xi},\bar{\xi}\right]$. For an Lk receiver, let $\Pi_{Rk}\left(a,\hat{\xi}\right)$ denote the expected payoff given the sender's message $\hat{\xi}$ and the receiver's action a; let $\mathbf{E}_k\left[\xi \mid \hat{\xi}\right]$ denote his updated belief about the expected value of ξ given message $\hat{\xi}$, and let $a_k\left(\hat{\xi}\right)$ denote his best response to the sender's message $\hat{\xi}$.

The $L0$ sender is naively truth-revealing and, hence, $\hat{\xi}_0(\xi) = \xi$. The $L0$ receiver naively believes that $\xi = \hat{\xi}$, i.e., whatever $L0$ sender reports. The $L0$ receiver's expected payoff is

$$\Pi_{R0}\left(a,\hat{\xi}\right) = r\mathbf{E}_0\left[\xi \mid \hat{\xi}\right]a - \frac{1}{2}ca^2 = r\hat{\xi}a - \frac{1}{2}ca^2. \tag{8}$$

An $L1$ sender believes the receiver is $L0$ and takes action $a_0\left(\hat{\xi}\right)$ in response to a message $\hat{\xi}$. Hence, her expected payoff is

$$\Pi_{S1}\left(\hat{\xi},\xi\right) = s\mathbf{E}\left[q \mid \xi\right]a_0\left(\hat{\xi}\right) = s\xi a_0\left(\hat{\xi}\right), \tag{9}$$

The payoffs for higher levels are defined iteratively in a similar manner. An Lk sender best responding to an $L(k-1)$ receiver has expected payoff

$$\Pi_{Sk}\left(\hat{\xi},\xi\right) = s\xi a_{k-1}\left(\hat{\xi}\right). \tag{10}$$

An Lk receiver best-responding to an Lk sender has expected payoff

$$\Pi_{Rk}\left(a,\hat{\xi}\right) = r\mathbf{E}_k\left[\xi \mid \hat{\xi}\right]a - \frac{1}{2}ca^2. \tag{11}$$

In particular, following a message $\hat{\xi} \in \Xi_k$ (i.e., $\hat{\xi}$ is in the Lk sender's feasible set of messages), the Lk receiver follows Bayesian inference to update her belief, $\mathbf{E}_k\left[\xi \mid \hat{\xi}\right] = \mathbf{E}\left[\xi \mid \xi \in \hat{\xi}_k^{-1}\left(\hat{\xi}\right)\right]$. If $\hat{\xi} \notin \Xi_k$, then the message cannot be from an Lk sender. In this case, following Kawagoe and Takizawa (2009) and Ellingsen and Östling (2010), we assume that an Lk receiver believes that the message is from the next highest sender type (lower than Lk) that uses this message.[6,7] Formally,

[6] Such a lower sender type always exists because the set of $L0$ sender's feasible messages is $[\underline{\xi},\bar{\xi}]$.

[7] Cai and Wang (2006) and Wang et al. (2010) alternatively assume that the receiver treats an off-equilibrium message as a mistake, and takes an action corresponding to the nearest on-equilibrium message. In our setting with large range of sender messages, their approach requires that the receiver ignores fairly large off-equilibrium deviations and further assumptions are required to fix the behavior of higher level players. Moreover, their approach is sensitive to the assumption that an Lk player assigns strictly zero probability to other players being of any type lower than $L(k\text{-}1)$. Consequently, the set of behaviors predicted by their approach will in general change if Lk player's belief is

$\mathbf{E}_k\left[\xi \mid \hat{\xi}\right] = \mathbf{E}\left[\xi \mid \xi \in \hat{\xi}_{k'}^{-1}\left(\hat{\xi}\right)\right]$ where k' is the highest sender type less than k for which $\hat{\xi} \in \Xi_{k'}$. We remark that this belief updating approach has intuitive appeal because it is equivalent to perturbing the Lk receiver's belief about the sender's type slightly such that there is a positive, albeit, arbitrarily small probability that the sender is of a lower type; sequential rationality and Bayesian inference then implies that the receiver believes that the sender is of the appropriate lower type.[8] Later, in §§6.1 and 6.3, we consider variations of the level-k model where there are no "unexpected messages" and the beliefs following any message are well-defined.

We solve for the behaviors of each player type iteratively, starting with the $L0$ receiver. The following theorem summarizes our results.

Theorem 3. *Under the level-k model, an Lk sender's strategy for $k > 0$ is $\hat{\xi}_k(\xi) = \hat{\xi}_k = \max\left\{\bar{\xi} - (k-1), \frac{\bar{\xi}+\underline{\xi}}{2}\right\}$. An Lk receiver's strategy for $k \geq 0$ is*

$$a_k\left(\hat{\xi}\right) = \begin{cases} a_I\left(\hat{\xi}\right) = \frac{r}{c}\hat{\xi}, & \hat{\xi} \leq \tilde{\xi}_k; \\ a_{NI} = \frac{r}{c}\frac{(\bar{\xi}+\underline{\xi})}{2}, & \textit{otherwise}, \end{cases}$$

where $\tilde{\xi}_k = \hat{\xi}_k - 1$.

Theorem 3 shows that an Lk sender ($k > 0$) distorts the message to a particular level that depends on her level of thinking, i.e., $\hat{\xi}_k(\xi) = \hat{\xi}_k$ (with some abuse of notation) as described in the theorem. For example, $L1$ distorts the message to $\hat{\xi}_1 = \bar{\xi}$ (the maximum message level), $L2$ sender distorts to $\hat{\xi}_2 = \bar{\xi} - 1$, and so on (recall that $\xi, \hat{\xi}$ are integers in the range $\left[\underline{\xi}, \bar{\xi}\right]$). Correspondingly, a higher level receiver is influenced by messages up to a threshold message level ($\tilde{\xi}_k = \hat{\xi}_k - 1$) that depends on his level of thinking, and ignores messages higher than this threshold. For example, an $L1$ receiver believes all messages $\hat{\xi} \leq \tilde{\xi}_1 = \bar{\xi} - 1$ and takes the fully-believing action $a_I\left(\hat{\xi}\right)$ following these messages, and ignores the message $\hat{\xi} = \bar{\xi}$ and takes the fully-disbelieving action a_{NI} for this message; an $L2$ receiver believes all messages $\hat{\xi} \leq \tilde{\xi}_2 = \bar{\xi} - 2$, taking action $a_I\left(\hat{\xi}\right)$, and ignores messages $\hat{\xi} > \tilde{\xi}_2$, taking action a_{NI}, and so on.

Intuitively, a strategic sender (higher than $L0$), being insatiable, sends the message that she thinks will induce the highest action from the receiver, namely, the highest message that she thinks the receiver will believe. Anticipating such sender behavior, a strategic receiver ignores messages

even slightly perturbed to allow for arbitrarily small positive probability of lower type players.

[8] Formally, the Lk receiver's behavior is consistent with the following prior belief about the sender's type: the sender is $L(k-1)$ with probability ϵ_1, is $L(k-2)$ with probability $\epsilon_1\epsilon_2$, is $L(k-3)$ with probability $\epsilon_1\epsilon_2\epsilon_3$ and so on, where $\epsilon_1, \epsilon_2, \epsilon_3 ...$ are positive and arbitrarily small. Therefore, conditional on receiving an off-equilibrium message, the receiver must believe (with probability arbitrarily close to 1) that the sender is of the highest type lower than Lk (and greater than $L0$) that uses this message; if no type between $L1$ and $L(k-1)$ uses this message, then the receiver must believe (with probability 1) that the sender is of type $L0$ since she uses all messages in $\left[\underline{\xi}, \bar{\xi}\right]$.

above a threshold message level. Boundedly-strategic thinking determines the sender's belief regarding which messages will be believed, and the receiver's belief about which messages have been distorted. For example, the $L1$ sender distorts the message to $\hat{\xi} = \overline{\xi}$, since she is insatiable and expects the $L0$ receiver to credulously believes all messages. An $L1$ receiver, anticipating the $L1$ sender's insatiable behavior, ignores the message $\hat{\xi} = \overline{\xi}$, expecting it to be uninformative. Any other message is taken to be from the next highest sender who could send that message, namely the $L0$ sender, and hence believed to be truthful. Proceeding iteratively, an Lk sender sends the message $\hat{\xi}_k$ that induces the highest action by an $L(k-1)$ receiver. Correspondingly, an Lk receiver believes all messages that are lower than the one sent by the Lk sender (i.e., $\hat{\xi} \leq \tilde{\xi}_k = \hat{\xi}_k - 1$) to be truthful (sent by a truth-revealing $L0$ sender), and ignores all other messages, believing them to be from an insatiable sender higher than $L0$ (message $\hat{\xi}_{k'} > \tilde{\xi}_k$ is taken to be from an Lk' sender, where $k' \in \{1,2...\ k\}$) and, hence, uninformative.

Thus, the level-k model also predicts distorted, yet informative and influential communication as well as heterogeneity in sender and receiver behaviors across the levels of thinking. Specifically, communication is informative because the $L0$ sender is truth-revealing; higher-level senders distort the message to a threshold level based on their level of thinking. Further, receivers are influenced by cheap-talk communication upto a threshold message level depending on their level of thinking.

3.4 Comparing the Predictions of the Behavioral Models

At this juncture, it is useful to compare the predictions from the trust-embedded model and the level-k model. Both models predict distorted yet informative and influential communication, as well as heterogeneity in sender and receiver behaviors. Yet, the underlying decision-making processes represented in these models are quite distinct. In particular, the sender being insatiable results in distinct predictions from the two models that can allow us to empirically distinguish between their predictions, especially, in games with relatively large strategy spaces such as the one we study. We now describe in what ways the predictions of the two models are similar and different.

For senders, both models allow for truthful behaviors. In the trust-embedded model, the sender is fully-trustworthy and truthful if her lying cost is sufficiently high, such that $A(\gamma, \bar{\alpha}_S) = 1$. For example, under the quadratic lying cost function $G\left(\left|\hat{\xi} - \xi\right|; \gamma\right) = \frac{\gamma}{2}\left(\hat{\xi} - \xi\right)^2$ that we use in the empirical application, the sender's trustworthiness factor $A(\gamma, \bar{\alpha}_S) = 1 + \frac{1}{\gamma} s \frac{r}{c}\left(\frac{1+\bar{\alpha}_S}{2}\right)$. For $\gamma \to \infty$, $A(\gamma, \bar{\alpha}_S) \to 1$. In the level-k model, the $L0$ sender is naively truth-revealing. Both models also allow for the sender who always distorts the message to the maximum extent possible, i.e., $\hat{\xi}(\xi) = \bar{\xi}$. In the trust-embedded model, if the lying cost is sufficiently small such that $A(\gamma, \bar{\alpha}_S)\underline{\xi} \geq \bar{\xi}$,

then $\hat{\xi}(\xi;\gamma,\alpha_S) = \bar{\xi}$. For example, in the experimental context, $\underline{\xi} = 10$ and $\bar{\xi} = 80$. Therefore, $A(\gamma,\bar{\alpha}_S) \geq 8$ would result in the sender always distorting the message to $\bar{\xi}$. In the level-k model, the type $L1$ sender always distorts the message to $\bar{\xi}$.

However, the two models predict different behaviors for senders who are neither truth-revealing nor always distorting to the maximum extent possible. The trust-embedded model predicts senders who partially inflate their message; the extent to which they inflate the message being constrained by the actual information ξ, since their cost of lying is increasing in the size of the lie. In contrast, the level-k model predicts senders who distort message to a particular message level that only depends on their level of thinking; being insatiable, the extent to which they distort their message is constrained only by their belief about the highest message that a lower level receiver will believe.

For receivers, both models allow for receivers who believe the sender is truth-revealing. In the trust-embedded model, the receiver is fully-trusting if $\alpha_R = 1$ and believes $\xi = \hat{\xi}$. In the level-k model, the $L0$ receiver naively believes $\xi = \hat{\xi}$. In either case, the receiver's optimal action is the fully-believing action $a_I\left(\hat{\xi}\right)$. The two models, however, differ in their predictions for receivers who are not fully-believing. In the trust-embedded model, if $\alpha_R < 1$, then the receiver, following a trust-based inference rule, is partially or moderately influenced by all messages. In particular, this is so even for $\alpha_R = 0$; in this case, the receiver believes that $\xi \sim Unif\left[\underline{\xi},\hat{\xi}\right]$ and hence $\mathbf{E}\left[\xi \mid \hat{\xi},\alpha_R\right] = \frac{\hat{\xi}+\underline{\xi}}{2}$. In the level-k model, an Lk receiver, following Bayesian inference and iterative thinking, forms a strategic belief that messages above a threshold are highly distorted and, hence, uninformative; essentially, the receiver anticipates that a sender higher than $L0$ is insatiable and always sends the message that the sender expects will induce the highest receiver action. As a result, the Lk receiver fully believes and is influenced by all messages up to a threshold level, and fully disbelieves and is uninfluenced by all messages higher than that threshold.

It is worth noting that prior empirical applications of the level-k model to cheap-talk experiments (Cai and Wang 2006; Wang et al. 2010) are set in a context in which the sender is satiable: the sender's payoff is strictly increasing in the receiver's action up to a point and then strictly decreasing thereafter. As a result, the sender's payoff structure directly limits the sender's incentive to distort her message. In particular, the extent to which the sender distorts her message also depends on the actual information. In turn, receivers are partially influenced by messages that they believe are distorted. These predictions are qualitatively similar to that of the trust-embedded model. Thus, one requires a suitable context in which the two models can be reliably distinguished. Our theoretical analysis in this section shows that a cheap-talk game in which the sender is insatiable

and the sender and receiver strategy spaces are large provides such a context.

4 Experiment Data and Empirical Methodology

We describe the experiment data from Özer et al. (2018) and discuss preliminary evidence that supports either behavioral model. To more systematically determine the extent to which either model is supported by the data, and what aspects of behavior each model captures well, a more sophisticated approach is necessary. We describe our empirical methodology for model comparison.

4.1 Experiment Data

The cheap-talk experiments in Özer et al. (2018) use the following experimental parameters: $\xi \sim Unif\,[10, 80]$, $\epsilon \sim Unif\,[-10, 10]$, $a \in [0, 180]$, $s = \frac{1}{2}$, $r = 1$, $c = 1$. Therefore, $\Pi_S = \frac{1}{2}qa$, $\Pi_R = qa - \frac{1}{2}a^2$ in Equations (1) and (2); and $a_{NI} = 45$, $a_I(\xi) = \xi \leq 80$ in Equations (3) and (4). Data is available for five experimental sessions - three main sessions and two additional sessions. Each session consisted of 12 participants and 11 paid decision rounds.[9] The value of ξ (and the unobserved market shock ϵ) was varied randomly from one decision round to the next. To minimize repeated interaction effects, participants were paired randomly and anonymously in each decision round and never rematched. Further, participants were rotated between the sender and the receiver roles between rounds, ensuring familiarity with both roles. Thus, each participant completed five decisions in one role, which could be sender or receiver, and six decisions in the other.

In the two additional experimental sessions, senders were encouraged to engage in more analytical thinking. Specifically, before making their decision, participants in the sender role were required to calculate their own expected payoffs if the receiver made a "high" ($a = 70$) or "low" ($a = 20$) decision, and participants in the receiver role were informed of this procedure. We expect the manipulation to affect the underlying drivers of behaviors under each behavioral model as follows. Under the trust-embedded model, the manipulation is likely to lower trust and trustworthiness, since making the pecuniary payoff salient and encouraging analytical thinking can crowd out non-pecuniary motives (e.g., Cappelletti et al. 2011; Cornelissen et al. 2011; Rand et al. 2012; Schulz et al. 2014; Zaki and Mitchell 2013). Under the level-k model, the manipulation is likely to encourage participants to engage in higher levels of thinking.

4.2 Preliminary Evidence for the Behavioral Models

Preliminary analysis suggests that there is prima facie support for either behavioral model. First, as shown in Özer et al. (2018), there is over-communication. In the main experiment sessions, sender's messages are significantly correlated with their private information ($\rho = 0.66$, $p < 0.01$), and the

[9] Participants played two unpaid practice rounds (one round in each role) prior to the 11 paid decision rounds.

receiver's actions are significantly correlated to the sender's message ($\rho = 0.36, p < 0.01$).[10] Second, there is considerable heterogeneity in behaviors across senders and receivers. Özer et al. (2018) use a tertile split based on the deviation from truth-revealing and fully-believing behaviors for senders and receivers (i.e., respectively, average $\hat{\xi} - \xi$ and $a - a_I\left(\hat{\xi}\right)$) to show that there is considerable difference between upper- and lower-most tertiles of senders and receivers.

Third, examining individual-level behaviors, there is preliminary evidence to support the $L0$ specification of the level-k model: 8 of 36 participants in the main session report their actual (true) information in at least three of their decision rounds in the sender role, and 3 participants deviate less than 2 units on average from the credulous (fully-believing) action in the receiver role.[11] Further, consistent with the predictions for higher level types starting with this $L0$ specification, 10 participants in the sender role roughly appear to distort the message to a particular level (i.e., irrespective of the true information), with 5 participants distorting close to the maximum level $\bar{\xi} = 80$ (median message ≥ 70) and 5 participants distorting to an intermediate level (median message < 70); 11 participants in the receiver role appear to be highly influenced by sender messages up to a threshold message level and then heavily discount the messages.[12] There is also preliminary evidence of behaviors predicted by the trust-embedded model. In the sender role, 10 participants are either truthful or inflate the message by a small amount (median inflation $< 10\%$), 8 participants inflate the message substantially (median inflation $> 40\%$), and 9 participants inflate message by an intermediate amount. In the receiver role, 7 participants appear to highly trust messages with relatively high action-level to message-level ratio (median ratio > 0.9), 10 participants appear to place low trust (median ratio < 0.5) and 8 participants appear to place moderate trust.

Lastly, the shift in behaviors in the additional (experimental manipulation) sessions compared to the main sessions appears consistent with what would be predicted from either model (as discussed above in §4.1). Senders distort their messages more in the additional sessions than in the main sessions: average $\hat{\xi} - \xi$ is 11.8 in the main session and 16.6 in the additional sessions ($p < 0.01$ under Kruskal-Wallis test for the difference). Further, only 1 of 24 participants in the additional sessions is truth-revealing in 3 or more decision rounds. And, receivers are less influenced by sender messages in the additional sessions than in the main sessions: average $a - a_I(\xi)$ is 17.8 in the main session and 19.9 in the additional session ($p = 0.08$ under Kruskal-Wallis test for the difference).

[10] All p-values are two-sided unless mentioned otherwise.

[11] In the sender role, 2 participants were truth-revealing in all decision rounds, 2 in all but one decision round, 3 in all but two decision rounds, and 1 was truth-revealing in three decision rounds.

[12] Remaining participants in either role could not be classified unambiguously; hence, the total count does not add up to 36. The ad-hoc classifications for the level-k and trust-embedded models are based on summary statistics of participants' individual-level behaviors.

These observations are consistent with less trusting and trustworthy behaviors in the additional sessions under the trust-embedded model, and with an increase in level of thinking leading to higher message distortion (e.g., senders switching from $L0$ to $L1$ thinking) and lower message influence under the level-k model.

4.3 Structural Model Specification

While the preliminary evidence supports either model, a more sophisticated and systematic analysis is necessary to compare their explanatory powers, including what aspects of behaviors each model captures well. In particular, a more sophisticated approach is necessary because the predicted behaviors depend on unobserved latent types of participants and, at the same time, observed behaviors can be "noisy" once we allow for "reasonable" deviations from theoretically predicted behaviors. Accordingly, we develop a structural model with latent types to address these issues.

We follow the econometric approach used in prior research for structural estimation of level-k and other non-equilibrium models (e.g., Haruvy et al. 2001; Crawford and Iriberri 2007; Wang et al. 2010). Specifically, we use a logit random-utility formulation for a player's decision. A player's utility from a decision is the sum of the player's predicted payoff (under a particular model) and a logit shock. The player chooses the decision that yields the highest utility.[13,14] The logit shock can cause the player to deviate from his or her theoretically predicted decision. Nevertheless, the theoretically predicted decision has a higher likelihood of being chosen and, hence, being observed in the data.[15] For both models, we allow for a mixture of player types to capture the heterogeneous behaviors across participants. All model parameters, including the probabilities for the player type mixture, are estimated from the data using maximum likelihood.

Model Likelihood. We construct the likelihood functions for each role as follows. Let N indicate the set of participants. Let $P \in \{S, R\}$ denote the player's role as a sender (S) or as a receiver (R). Let Ω_P denote the set of player types for role P, and π_ω denote the probability of player type $\omega \in \Omega_P$ such that $\sum_{\omega \in \Omega_P} \pi_\omega = 1$. In the trust-embedded model, player type refers to

[13] An alternative approach is to estimate the model using a "predicted behavior plus noise" strategy, which simply adds errors from a specified distribution (e.g., truncated normal distribution) to the model predicted decisions. In this case, the deviations are determined only by the noise distribution and are not payoff sensitive. In contrast, the approach we adopt imposes theory-based structure also on the deviations.

[14] One exception is the $L0$ player in the level-k model, whose predicted behavior is not payoff-based. In this case, we adopt the "predicted behavior plus noise" approach as explained in Equation (13).

[15] Wang et al. (2010) find that in their experiments (with small strategy spaces), participants often chose a theoretically predicted action. Therefore, they adopt a "spiked logit" formulation in their model estimation: only the deviation from the predicted behavior follows a logit distribution, and the "spike" of the probability with which the predicted behavior is chosen is estimated from the data. As in Crawford and Iriberri (2007), in our context (with large strategy spaces), participants often do not choose the precise theoretically predicted action and we adopt the standard logit formulation.

the trust or trustworthiness types. In the level-k model, a player type refers to the level of thinking. Let T_{iP} denote the set of rounds in which participant i plays role P. Participant i in the sender's role in round $t \in T_{iS}$ observes state ξ_t and sends a message $\hat{\xi}_{it}$. Participant i in the receiver's role in round $t \in T_{iR}$ observes message $\hat{\xi}_{it}$ and takes action a_{it}. Define I_{it}, D_{it}, respectively, as the information observed and decision made by participant i in round t such that $\{I_{it}, D_{it}\} = \left\{\xi_t, \hat{\xi}_{it}\right\}$ for the sender role, and $\{I_{it}, D_{it}\} = \left\{\hat{\xi}_{it}, a_{it}\right\}$ for the receiver role. Let Δ_P denote the set of feasible decisions in role P; Δ_S is the set of integers in $\left[\underline{\xi}, \bar{\xi}\right]$ and Δ_R is the set of integers in $[0, \bar{a}]$.

Except for the $L0$ sender in the level-k model (whose theoretical behavior is not payoff-based), we can specify a player's payoff from choosing various actions including his or her theoretically predicted behavior. Let $\Pi_P\left(D_{it}, I_{it}; \theta_\omega\right)$ denote participant i's payoff in role P in round $t \in T_{iP}$ from decision D_{it} given information I_{it} if the participant's type is $\omega \in \Omega_P$, where θ_ω denotes type specific parameters in the payoff function. In the trust-embedded model, Π_P is given by Equations (5) and (7). In the level-k model, Π_P is given by Equations (10) and (11). Then, under the logit random-utility formulation, the probability that participant i in role P, conditional on being type $\omega \in \Omega_P$, makes decision $D_{it} \in \Delta_P$ given information I_{it} in round $t \in T_{iP}$ is

$$Pr_P(D_{it} \mid I_{it}, \theta_\omega, \lambda_\omega) = \frac{\exp\{\lambda_\omega \cdot \Pi_P(D_{it}, I_{it}; \theta_P(\omega))\}}{\sum_{D \in \Delta_P} \exp\{\lambda_\omega \cdot \Pi_P(D, I_{it}; \theta_P(\omega))\}}, \tag{12}$$

where λ_ω represents the precision of the logit errors; we allow for precision to be type-specific. Note that if $\lambda_\omega \to \infty$, then the player strictly maximizes his or her payoff to choose the theoretically predicted action. Whereas if $\lambda_\omega \to 0$, then the player uniformly randomizes across all actions.

In the case of the level-k $L0$ sender, following Wang et al. (2010), we consider that the sender deviates from her truth-revealing behavior $\hat{\xi} = \xi$ as per a truncated normal shock e with zero mean and variance σ_{L0}^2, such that the observed report $\hat{\xi} = \xi + e \in \left[\underline{\xi} - 0.5, \bar{\xi} + 0.5\right]$. Specifically, given true information ξ, the report $\hat{\xi}$ is the closest integer to $\xi + e$. Therefore, the likelihood that participant i in the sender role conditional on being type $L0$ chooses a report $\hat{\xi}_{it}$ given true information ξ_t in round $t \in T_{iS}$ is

$$Pr_S(\hat{\xi}_{it} \mid \xi_t, \sigma_{L0}) = \frac{\Phi\left(\frac{\hat{\xi}_{it} - \hat{\xi}_t + 0.5}{\sigma_{L0}}\right) - \Phi\left(\frac{\hat{\xi}_{it} - \hat{\xi}_t - 0.5}{\sigma_{L0}}\right)}{\Phi\left(\frac{\bar{\xi} - \hat{\xi}_t + 0.5}{\sigma_{L0}}\right) - \Phi\left(\frac{\underline{\xi} - \hat{\xi}_t - 0.5}{\sigma_{L0}}\right)}, \tag{13}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution.

The likelihood of observing decisions $\mathbf{D}_{iP} = \{D_{it} : t \in T_{iP}\}$ by participant i in role P given information $\mathbf{I}_{iP} = \{I_{it} : t \in T_{iP}\}$ unconditional on type is

$$L_{iP}(\mathbf{D}_{iP} \mid \mathbf{I}_{iP}, \pi^P, \lambda^P, \theta^P) = \sum_{\omega \in \Omega_P} \pi_\omega \prod_{t \in T_{iP}} Pr_P(D_{it} \mid I_{it}, \theta_\omega, \lambda_\omega), \tag{14}$$

where $\pi^P = \{\pi_\omega : \omega \in \Omega^P\}$ is the distribution of types, $\lambda^P = \{\lambda_\omega : \omega \in \Omega^P\}$ and $\theta^P = \{\theta_\omega : \omega \in \Omega^P\}$, respectively, are type-specific precision and payoff parameters. Hence, the model log-likelihood function for role P given observations $\mathbf{D}^P = \{\mathbf{D}_{iP} : i \in N\}$ and $\mathbf{I}^P = \{\mathbf{I}_{iP} : i \in N\}$ is

$$LL_P(\pi^P, \lambda^P, \theta^P | \mathbf{D}^P, \mathbf{I}^P) = \sum_{i \in N} \log\left(L_{iP}(\mathbf{D}_{iP} \mid \mathbf{I}_{iP}, \pi^P, \lambda^P)\right). \tag{15}$$

We estimate $\{\pi^P, \lambda^P, \theta^P\}$ that maximize the above log-likelihood for each role for each model.

Trust-Embedded Model Player Types. We allow for three trustworthiness types for senders and three trust types for receivers; the three types being designated as high (H), medium (M) and low (L) such that $\omega \in \{H, M, L\}$ for either role. Allowing for additional types does not improve model performance (per the log-likelihood ratio test). As discussed in Appendix C, for senders, the logit choice probabilities in the likelihood function (in Equation 12) are affected only by their trustworthiness factor $A(\gamma_\omega, \bar{\alpha}_{S\omega})$ and the scaled lying cost $\lambda_\omega \gamma_\omega$. As a result, only these type-specific parameters are estimable. For receivers, the trust type α_ω and the logit error precision λ_ω are estimable.

Level-k Model Player Types. We allow for player types up to $L3$ and also include an LH type whose level of thinking $H > 3$ is estimated from the data. Prior empirical applications of the level-k model across a wide variety of games have found that player types higher than $L3$ are rare, with most players being of type $L1$ or $L2$ (e.g., see surveys by Crawford et al. 2013; Georganas et al. 2015). We include an LH type since in our application context, the iterative logic needed to go from one level of thinking to the immediately higher level is relatively straightforward compared to many other games. For example, an $L1$ sender expects the receiver to believe messages $\xi \leq 80$, an $L2$ sender expects the receiver to believe messages $\hat{\xi} \leq 79$ and so on. So a player who anticipates $L1$ thinking could relatively easily anticipate $L2$ thinking, in effect thinking in larger steps than implied by the literal application of the level-k model. The LH type accounts for this possibility. Thus, we provide more flexibility to the level-k model than implied by its literal application.

Furthermore, we find that combining the lower level types in the level-k model into fewer types does not affect the model likelihood by much, and hence yields a more parsimonious model that has substantially better AIC and BIC statistics. Essentially, an Lk player's payoff function changes gradually with the level of thinking. For example, the payoff functions for $L1$ and $L2$ senders differ only for $\hat{\xi} = 80$, since the $L2$ sender believes that this message will be ignored (by an $L1$ receiver) whereas the $L1$ sender believes that this message will be fully believed (by an $L0$ receiver). Similarly, the payoff functions for $L2$ and $L3$ differ only for $\hat{\xi} = 79$ and so on. In the case of receivers, the

payoff functions for $L0$ and $L1$ receivers differ only for $\hat{\xi} = 80$, the payoff functions for $L1$ and $L2$ differ only for $\hat{\xi} = 79$ and so on. Therefore, to avoid unduly penalizing the level-k model for additional parameters when comparing with the trust-embedded model (using the AIC and BIC statistics), we combine the lower level player types. For senders, we combine $L1$ to $L3$ types into an $L1 \sim 3$ type that has the same payoff function as an $L1$ sender (which yields the best fit). For receivers, we combine $L0$ to $L3$ types into an $L0 \sim 3$ type with the same payoff function as an $L0$ receiver (which yields the best fit). We thus estimate three sender types ($\omega \in \{L0, L1 \sim 3, LH\}$), and two receiver types ($\omega \in \{L0 \sim 3, LH\}$).

4.4 Model Comparison Measures

We compare the models with respect to their in-sample fit and out-of-sample forecasting performance using the data from the main experimental sessions. We also compare their ability to recover the effect of the experimental manipulation in the additional sessions. For in-sample fit, we use the AIC and BIC of the estimated models; a lower AIC and BIC indicates a better model. For out-of-sample forecasting performance, we use the following measures: (i) *Mean Squared Error* (MSE) between the observed and predicted behaviors; a lower mean square error indicates better forecasting performance, (ii) *Goodness of fit* of predictions, where we regress predicted behavior with observed behavior and compare regression slope $\hat{\beta}$ and regression R^2; a higher $\hat{\beta}$ and a higher R^2 indicate better goodness of fit. To compare the ability to recover the effect of the experimental manipulation, we evaluate whether the change in estimated model parameters between the main and additional sessions is consistent with what each behavioral theory would predict (as explained in §4.1).

To reliably determine out-of-sample forecasting performance with limited data, we use Monte Carlo Cross-Sample Validation (MCCV) (Hastie et al. 2011). Under MCCV, the data is randomly partitioned into a training sub-sample, and a validation sub-sample. Each model is estimated on the training sub-sample. Next, each participant's sender and receiver types are determined based on their estimated posterior probability of being a particular type. Then, their behavior in the validation sub-sample is predicted and compared with their observed behaviors. Importantly, to minimize the risk of over-fitting or bias in the selection of the sub-samples that is possible with limited data, we repeat the above process 1000 times and calculate the average forecasting performance across these iterations. To partition the data in each iteration, we randomly choose 4 decisions of each player in each role in the training sub-sample, leaving out the participant's remaining (1 or 2) decisions in that role to be in the validation sub-sample.

5 Which Model Explains Behaviors Better?

5.1 Model Comparison for Senders

Table 1 summarizes the estimation results for senders.[16] For either model, the $\lambda \cdot \gamma$ and λ estimates being significantly different than zero indicate that the fitted behaviors are driven by the underlying theoretical model predictions (and not just the logit error).[17] For the trust-embedded model, we observe that the high- and medium-trustworthy types are relatively truthful, with trustworthiness factors $A_H = 1.02$ and $A_M = 1.07$, respectively; the low-trustworthy type inflates messages considerably, with a trustworthiness factor $A_L = 2.5$, though not to the maximum extent (which would require $A_L = 8$, such that $A_L\xi = \bar{\xi}$ for $\xi = 10$ and $\bar{\xi} = 80$). A substantial number of participants are classified as medium- and low- trustworthy types, and a smaller fraction as being of high-trustworthy type.[18] For the level-k model, we observe that participants are mainly classified as type $L0$, being truth-revealing, or of type $L1 \sim 3$, always distorting the messages to the maximum extent possible (to $\bar{\xi} = 80$). For $L0$ participants, $\sigma_{L0} = 6.2$ indicates some participants classified as $L0$ may be nevertheless distorting messages by a non-neglibible extent. Only one participant is classified as type LH, distorting to a message level $\hat{\xi}_{LH} = 48$.

Comparing the in-sample model fits of both models, the trust-embedded model performs better, with lower AIC and BIC. The out-of-sample forecasting accuracy is also better for the trust-embedded model, with considerably lower prediction errors, and higher $\hat{\beta}$ and R^2. Finally, both models appear to directionally recover the effect of the experimental manipulation as the change in the estimated model parameters is consistent with the respective behavioral theory prediction (discussed in §4.2). Specifically, the trust-embedded model shows a higher proportion of low-trustworthy type in the additional sessions than in the main sessions (70.83% vs. 55.56%, $p = 0.12$), with the trustworthiness of the low-trustworthy type being lower than before (higher A_L than in the main sessions), both of which translate to higher inflation in sender messages. Similarly, there are fewer high-trustworthy types in the additional sessions and the trustworthiness of the medium-trustworthy type is lower (higher A_M than in the main sessions). Under the level-k model, there is an increase in the proportion of $L1 \sim 3$ and in the proportion of LH types at the expense of the $L0$ type

[16] In all model estimate tables, * indicates parameter is significantly different than zero at 2.5% confidence level. We use bootstrapping to determine significance level (since the models have latent classes).

[17] The low $\lambda_L \cdot \gamma_L$ and high $\lambda_H \cdot \gamma_H$ estimates simply indicate low γ_L and high γ_H, consistent with the A_L and A_H estimates.

[18] Participants are classified to be of the type for which their estimated posterior probability of being a particular type is highest. The model-based classifications, in general, need not match the ad-hoc "model-free" classifications (in §4.2) because the model-based classifications account for reasonable deviations from theoretically predicted behaviors and classifies participants into limited number of player types (based on the likelihood criterion) to avoid over-fitting.

(47.22% vs. 29.17%, $p = 0.08$) in the additional sessions, which is consistent with sender participants engaging in higher level of thinking as predicted by the level-k theory. This shift in the level of thinking should cause sender messages to be more inflated in the additional sessions. Nevertheless, the trust-embedded model performs better in explaining the observed behaviors in the additional sessions, with lower AIC and BIC.

Table 1: Level-k and Trust-Embedded Model Estimation Results for Senders

		Trust-embedded Model		Level-K Model	
Model Estimates	Classification	$High$	4 (11.11%)	$L0$	17 (47.22%)
		$Medium$	12 (33.33%)	$L1 \sim 3$	18 (50.00%)
		Low	20 (55.56%)	LH	1 (2.78%)
	Model Parameters	A_H	1.02*	$\hat{\xi}_{LH}$	48*
		A_M	1.07*	σ_{L0}	6.20*
		A_L	2.50*	$\lambda_{L1\sim3}$	1.34*
		$\lambda_H \cdot \gamma_H$	809.49*	λ_H	16.67*
		$\lambda_M \cdot \gamma_M$	24.74*		
		$\lambda_L \cdot \gamma_L$	0.84*		
In-Sample Model Fit	LL	-693.03		-711.58	
	AIC	1402.07		1435.15	
	BIC	1428.37		1454.88	
Out-of-Sample Performance	MSE	305.82		355.06	
	$\hat{\beta}$	0.79		0.69	
	R^2	0.49		0.39	
Experimental Manipulation	Classification	$High$	1 (4.17%)	$L0$	7 (29.17%)
		$Medium$	6 (25.00%)	$L1 \sim 3$	12 (50.00%)
		Low	17 (70.83%)	LH	5 (20.83%)
	Model Parameters	A_H	1.03*	$\hat{\xi}_{LH}$	47*
		A_M	1.21*	σ_{L0}	9.09*
		A_L	2.99*	$\lambda_{L1\sim3}$	2.38*
		$\lambda_H \cdot \gamma_H$	612.74*	λ_H	2.34*
		$\lambda_M \cdot \gamma_M$	25.60*		
		$\lambda_L \cdot \gamma_L$	0.84*		
	LL	-466.68		-481.62	
	AIC	949.36		975.23	
	BIC	972.42		992.53	

5.2 Model Comparison for Receivers

Table 2 summarizes the estimation results for receivers. For the trust-embedded model, we observe that the high-trust type is highly influenced by sender messages ($\alpha_{RH} = 0.72$, not significantly different than 1), while the medium- and low-trust types are influenced much less and substantially

discount sender messages ($\alpha_{RM} = 0.13$ and $\alpha_{RL} = 0$).[19] A majority of participants are classified as low-trust type, and a substantial proportion are classified to be of high-trust type. Under the level-k model, most participants are classified to be of LH type, believing messages only up to a message level $\tilde{\xi}_{LH} = 46$ and disbelieving all higher messages. A small fraction are classified as $L0 \sim 3$ that are highly influenced by sender messages; however, $\lambda_{L0\sim3}$ is not significantly non-zero, which suggests that the behaviors of these participants is not captured well by the level-k model.

Table 2: Level-k and Trust-Embedded Model Estimation Results for Receivers

		Trust-embedded Model		Level-K Model	
Model Estimates	Classification	$High$	11 (30.56%)	$L0 \sim 3$	6 (16.67%)
		$Medium$	3 (8.33%)	LH	30 (83.33%)
		Low	22 (61.11%)		
	Model Parameters	α_{RH}	0.72*	$\tilde{\xi}_{LH}$	46*
		α_{RM}	0.13	$\lambda_{L0\sim3}$	23.49
		α_{RL}	0.00	λ_{LH}	4.27*
		λ_H	24.76*		
		λ_M	122.25*		
		λ_L	3.67*		
In-Sample Model Fit	LL	-749.67		-779.66	
	AIC	1515.33		1567.32	
	BIC	1541.64		1580.47	
Out-of-Sample Performance	MSE	174.18		194.45	
	$\hat{\beta}$	0.56		0.47	
	R^2	0.41		0.38	
Experimental Manipulation	Classification	$High$	4 (16.67%)	$L0 \sim 3$	2 (8.33%)
		$Medium$	7 (29.17%)	LH	22 (91.67%)
		Low	13 (54.17%)		
	Model Parameters	α_{RH}	0.62*	$\tilde{\xi}_{LH}$	50*
		α_{RM}	0.36*	$\lambda_{L0\sim3}$	8.66
		α_{RL}	0.00	λ_{LH}	4.47*
		λ_H	77.32*		
		λ_M	6.97*		
		λ_L	4.73*		
	LL	-501.78		-525.87	
	AIC	1019.55		1059.74	
	BIC	1042.61		1071.27	

Comparing the in-sample model fits of both models, the trust-embedded model performs considerably better; its AIC and BIC are substantially lower. Further, the trust-embedded model also outperforms the level-k model in out-of-sample forecasting accuracy, with lower prediction error

[19] Low trust receivers are still influenced by the sender's message $\hat{\xi}$ because they believe that the true state is less than the received message. Specifically, from Theorem 2 in §3.2, $a^*\left(\hat{\xi}; \alpha_R = 0\right) = \frac{r}{c}\mathbf{E}\left[\xi \mid \xi < \hat{\xi}\right] = \frac{1}{2}\left(\hat{\xi} + \underline{\xi}\right)$.

and modestly higher $\hat{\beta}$ and R^2. Finally, both models appear to directionally recover the effect of the experimental manipulation. Specifically, the trust-embedded model finds a lower proportion of high-trust type in the additional sessions than in the main sessions (16.67% vs. 30.56%, $p = 0.11$). The level-k model finds there are fewer $L0$ players (8.33% vs. 16.67%, $p = 0.18$). Nevertheless, the trust-embedded model has better in-sample fit than the level-k model in the additional sessions, with considerably lower AIC and BIC.

5.3 Why Does the Trust-Embedded Model Perform Better?

To understand what aspects of behaviors the trust-embedded model captures better than the level-k model, it is useful to decompose the in-sample prediction error (MSE) for each model by participant types. In the case of senders, note that the low-trustworthy type in the trust-embedded model (20 participants) and the $L1 \sim 3$ and LH types (19 participants, 18 are $L1 \sim 3$, 1 is LH) in the level-k model capture senders who significantly distort their messages. We find that the trust-embedded model captures their behaviors better: in-sample MSE for the low-trustworthy type is 460.53 and for the $L1 \sim 3$ and LH types is 647.87 (676.18 for $L1 \sim 3$ participants, and 138.33 for LH participant). The trust-embedded model also performs better for senders who do not distort much, albeit by a smaller margin: the in-sample MSE for high- and medium-trustworthy types is 26.07 (1.18 for 4 high-trustworthy participants and 34.37 for 12 medium-trustworthy participants), whereas it is 35.49 for the $L0$ type in the level-k model (17 participants). In this case, the trust-embedded model better describes behaviors that are highly-trustworthy but not fully truthful through the medium-trustworthy type. These differences between the two models in explaining behaviors is also highlighted by the figures in Appendix D, which depict the predicted vs. observed behaviors for different sender types under either model. Thus, the the trust-embedded model performs better because the behaviors uniquely predicted by it are relatively more important in explaining observed sender behaviors; namely, that of senders who are neither truth-revealing nor always distorting to the maximum extent.

Similarly, the behaviors uniquely predicted by the trust-embedded model for receivers, namely those who are not fully-believing, are relatively more important in explaining observed receiver behaviors. The in-sample MSE for medium- and low-trust types in the trust-embedded model is 189.48 (8.23 for 3 medium-trust participants and 214.2 for 22 low-trust type participants), whereas it is 199.18 for the LH type in the level-k model (30 participants). The in-sample MSE for participants classified as being highly influenced is relatively similar for the two models: 40.29 for the high-trust type in the trust-embedded model, and 38.63 for the $L0 \sim 3$ type in the level-k model. However,

more participants are classified as being highly influenced under the trust-embedded model than in the level-k model, which further contributes to the better performance of the trust-embedded model across participants. The figures in Appendix D highlight these differences in explanatory powers by depicting the predicted vs. observed behaviors for different receiver types under either model.

6 Level-k Model Variations

Our results above indicate that the trust-embedded model has better explanatory power than the level-k model. In this section, we investigate whether the explanatory power of the level-k model is improved by relaxing certain assumptions.

6.1 Player's Belief About Others' Level of Thinking

The cognitive hierarchy (CH) model differs from the level-k model in that a higher-level player assigns a positive probability to others being of any lower level type up to $L0$. Consequently, an Lk receiver always expects to receive messages over the entire support $\left[\underline{\xi}, \bar{\xi}\right]$, since he assigns positive probability to the sender being $L0$. Hence his beliefs following any message is well-defined without requiring further assumptions. We defer the details of our analysis to Appendix E. We find that, in the context of cheap talk by an insatiable sender, the CH model does not predict substantially new behaviors and, hence, does not help address the gaps in the level-k model's explanatory power. Specifically, the predicted behavior of an Lk player in the CH model is either exactly the same (for sender) or almost the same (for receiver) as an Lk or a lower level player in the level-k model. Intuitively, an Lk sender, being insatiable, always distorts the message to a particular level, namely the level that she believes will induce the highest expected action from the receiver (given the distribution of lower level receiver types). In turn, an Lk receiver holds the strategic belief that messages above a threshold are highly distorted; while there is a small probability that the message could be from a truthful $L0$ sender, there is a much higher probability that the message is from a higher sender type that is distorting its message. Hence, messages above a threshold are heavily discounted. Thus, no substantially new behaviors are predicted by the CH model. In fact, the CH model predicts a narrower range of behaviors across player types than the level-k model. For example, higher level senders may not distort messages to a level below $\hat{\xi} = 76$ because they assign low probability to higher level receivers (see Appendix E).

6.2 Mixture of $L0$ players

In applications of the level-k model to complete information games, naive behavior of $L0$ players is typically specified as uniformly randomizing across all actions. A natural question is whether such a specification can improve the performance of the level-k model in a cheap-talk application

(a game of incomplete information). Note that if all $L0$ players are randomizing, then the level-k model cannot explain informative communication. Specifically, an $L1$ receiver best-responding to a randomizing $L0$ sender will ignore all messages; an $L1$ sender best-responding to a randomizing $L0$ receiver is indifferent between all messages and, hence, may simply randomize. Therefore, we instead allow for a mixture of $L0$ senders and receivers, where a fraction $\eta \in (0,1)$ of $L0$ senders are truthful and a fraction $1-\eta$ are uniformly randomizing (irrespective of ξ); also, a fraction $\mu \in (0,1)$ of $L0$ receivers are credulous and a fraction $1-\mu$ are uniformly randomizing (irrespective of $\hat{\xi}$). We defer the details of our analysis to Appendix F.

We find that allowing for a mixture of $L0$ players mainly affects the behavior of higher-level receivers, but not of higher-level senders. As before, a higher level sender, being insatiable, distorts to a particular message level depending on her level of thinking. A higher level receiver fully disbelieves all messages above a threshold. However, below this threshold, the receiver does not fully believe the message; he believes the message is from a truth-revealing $L0$ sender with probability η and from a randomizing $L0$ sender with probability $1-\eta$. Hence, the receiver is only partially influenced. This behavior is somewhat similar to the receiver in the trust-embedded model though not exactly the same; in particular, unlike in the trust-embedded model, $\mathbf{E}\left[\xi \mid \hat{\xi}\right] > \hat{\xi}$ for $\hat{\xi}$ low enough for an Lk receiver.[20]

We find that this level-k model, in general, does not perform better than the original level-k model; it provides modest improvement over the original level-k model with regards to AIC in the main experimental sessions, but performs worse on all other measures. Table 5 in Appendix F shows the estimation results. As such, our results support the simpler $L0$ assumptions used in previous work for cheap-talk games.

6.3 Trembling Behavior

We examine the implication if the Lk sender may "tremble" while sending her message, deviating to messages other than $\hat{\xi}_k$, such that all messages are expected from an Lk sender by the Lk receiver. Formally, an Lk player (sender and receiver) makes decisions following a random-utility choice process with *idiosyncratic* logit shocks of precision λ to the *systematic* payoffs in Equations (1) and (2), and anticipates that others also behave in this way (i.e., others also experience logit

[20] Conditional on the message being from an $L0$ sender who is not truthful, $\mathbf{E}\left[\xi \mid \hat{\xi}\right] = 45$, which is higher than $\hat{\xi}$ for $\hat{\xi} < 45$. Therefore, when there is a mixture of $L0$ senders, $\mathbf{E}\left[\xi \mid \hat{\xi}\right] > \hat{\xi}$ for $\hat{\xi}$ low enough. In the trust-embedded model, conditional on the message not being truthful, the message is believed to be inflated and the true information could be any value less than the message (consistent with a sender with a low enough lying cost). Consequently, $\mathbf{E}\left[\xi \mid \hat{\xi}\right] < \hat{\xi}$ always.

shocks to their systematic payoffs). This model may also be seen as incorporating elements of the AQRE model within the level-k model. The model is solved numerically. We find that the $L1$ sender messages can be partially informative, albeit to a limited extent. Essentially, the magnitude of the sender's systematic payoff is higher if ξ is higher. Hence, as shown in Appendix G, the $L1$ sender is less likely to tremble from distorting the message to $\bar{\xi}$ (and send lower messages) if ξ is higher, i.e., higher messages are more likely if the true information is higher; as a result, $\hat{\xi}$ is partially informative. In turn, the $L1$ receiver is partially influenced by the $L1$ sender's messages (instead of being either fully-believing or fully-disbelieving). Figure 3 in Appendix G shows examples of predicted behaviors for specific values of λ. We observe that the behavior of each higher level in this level-k model is quite distinct from the immediately lower level. Moreover, each higher level of thinking is no longer a trivial extension of the lower level. Thus, we do not expect players to think in "larger steps". Accordingly, in the empirical estimation, we consider three distinct player types, namely $L0$, $L1$ and $L2$. Table 6 in Appendix G provides the estimation results and Figure 4 in the same appendix illustrates the estimated behaviors. We find that the model performance is worse for senders and about the same for receivers compared to the original level-k model.

7 Hybrid Models

7.1 Including Level-k Behaviors in the Trust-Embedded Model

While the trust-embedded model performs better overall in capturing behaviors across all participants taken together, it is possible that the behaviors of some participants are better represented by the level-k model. We now examine whether the trust-embedded model can be improved upon by including the behaviors uniquely predicted by the level-k model, namely the LH sender type who distorts to a particular message level but not to the maximum extent, and the LH receiver type who is fully-believing up to a threshold message level and ignoring messages above this threshold.

We find that this hybrid model performs consistently better than the trust-embedded model in the case of receivers; for senders, the hybrid model performs modestly better or worse depending on the measure. Table 7 in Appendix H provides the estimation results. For senders, 1 participant in the main experimental sessions and 5 participants in the additional sessions are classified as LH type; these are mainly participants classified as low trustworthiness senders in the original trust-embedded model. For receivers, 5 participants in the main experimental sessions and 3 participants in the additional session are classified as LH type; mainly those classified as being of low- or medium-trust in the original trust-embedded model (recall that medium-trust receivers had relatively low levels of trust). In all cases, the estimated behaviors of the LH types differs substantially from

that of the trust and trustworthy types in the trust-embedded model. For example, in the main experimental sessions, the LH sender distorts messages to $\hat{\xi}_{LH} = 48$, and the LH receiver is fully-believing of messages up to $\tilde{\xi}_{LH} = 69$ and is then fully-disbelieving of higher messages.

Overall, our results indicates the presence of both types of decision processes across individuals. We further observe that the behaviors uniquely predicted by the trust-embedded model (i.e., senders that partially inflate messages, receivers who are partially influenced) better explain behaviors of more participants. In this sense, the trust-embedded model emerges as the dominant explanation.

7.2 Embedding Lying Cost in the Level-k Model

We investigate whether the process by which a given individual arrives at his or her decisions includes elements of both behavioral theories. Specifically, we introduce lying cost for senders from the trust-embedded model to the level-k model. We defer the analysis details to Appendix I. In the presence of lying cost, the $L1$ sender's messages can be partially informative, similar to the trust-embedded model; consequently, the $L1$ receiver is influenced by the $L1$ sender's messages, albeit partially since the receiver anticipates the messages are distorted. In turn, the $L2$ sender's message will be tailored to influence the $L1$ receiver in the presence of lying costs and can be informative, and so on. Moreover, the behavior of each higher level is quite distinct from the immediately lower level, and is not a trivial extension of a lower level of thinking. Therefore, in the empirical estimation, we consider three distinct player types, namely $L0$, $L1$ and $L2$. We allow for $L1$ and $L2$ sender types to differ in their lying cost parameter γ, and $L1$ and $L2$ receivers to differ in their beliefs about the sender's lying cost.

We find that incorporating lying cost in the level-k model can improve the performance of the level-k model substantially. However, the trust-embedded model on its own still performs better and represents a better parsimony-to-explanatory power trade-off. Table 8 in Appendix I provides the estimation results. The in-sample performance of the hybrid model for senders is very close to that of the trust-embedded model. In fact, the estimated behaviors are quite close to the behaviors of the corresponding trustworthy types, with $L0$, $L1$ and $L2$ roughly matching high, low and medium trustworthy types, respectively. In this sense, the level-k model converges to the trust-embedded model for senders and does not display characteristics peculiar to the level-k model (which can occur for other parameter values). The out-of-sample performance is however much worse than the trust-embedded model (recall that under MCCV, the model is re-estimated in each new sub-sample). For receivers, we find that the in-sample performance is better than the original level-k model. Essentially, the ability of the model to allow for receivers to be partially influenced by

sender messages represents an improvement over the original level-k model. Further, the estimated behaviors do display some characteristics peculiar to the level-k model; for instance, the $L2$ receiver discounts messages differently depending on whether it believes the messages is from the $L2$, $L1$ or $L0$ receiver. Nevertheless, its performance still falls short of the trust-embedded model. Figure 5 in Appendix I illustrates the predicted behaviors for the estimated sender and receiver types.

8 Summary and Discussion

In this paper, we provide the first direct comparison of two leading behavioral economics theories of cheap talk that are based on fundamentally contrasting perspectives of human behaviors. We identify a cheap-talk context that well represents many practical business and economic situations, namely cheap talk by an insatiable sender, and show that the two theories lead to characteristically distinct and empirically distinguishable predictions in this setting. Leveraging past experiment data from this setting, we compare the explanatory powers of both theories to shed light on what aspects of behaviors each theory captures well, which theory on its own describes behavior the best, and whether a combined model can improve the explanatory power.

Overall, we find that cheap talk behaviors are better explained by the more positive perspective of human behavior, namely, that decision makers are guided by non-pecuniary motives to be trusting and trustworthy. Thus, effective cheap talk is not simply a matter of boundedly-strategic thinking in the sole pursuit of self-interest. In a cheap-talk experiment with an insatiable sender, the behaviors uniquely predicted by the trust-embedded model explain sender as well as receiver behaviors of most participants better than the level-k model. In particular, senders appear to be constrained in how they distort messages by non-pecuniary lying costs, and not just motivated by their pecuniary incentive to be insatiable. Also, receivers appear to be influenced by messages based on their intrinsic tendency to trust, and not based on Bayesian inference and iterative strategic thinking. While there is evidence of boundedly-strategic thinking for some individuals, trust and trustworthiness emerges as the dominant explanation.

Our results lead to implications for managers as well as scholars studying cheap talk. From a managerial perspective, even in cheap-talk situations with significant difference of interests between parties (as is the case with cheap talk by an insatiable sender), we find trust and trustworthiness play a prominent role. Hence, firms should focus on designing processes to reduce barriers for and to engineer trusting and trustworthy relationships; for example, by appointing a fixed contact person and avoiding management rotations (Özer et al. 2014; Li et al. 2019), making relationship-specific investments (Beer et al. 2018), or facilitating interactive video communication between

participants on an online platform (Özer and Zheng 2019), reducing perceived vulnerabilities and social uncertainties in market environment (Donohue et al. 2020), or through suitable reputation and feedback systems (Bolton et al. 2013). From a research perspective, our results indicate that for modeling cheap-talk interactions, the trust-embedded model provides the best parsimony-to-explanatory power trade-off. For researchers wanting to use the level-k model, we would suggest that incorporating sender lying cost from the trust-embedded model can significantly improve the explanatory power of the level-k model.

Our work also provides insights about developing a common context to compare models of boundedly-strategic behavior and non-pecuniary motives. As noted by Sobel (2013), distinguishing between these theories is an an important gap in the current understanding of cheap talk. However, doing so can be challenging because both theories can lead to qualitatively similar predictions. Our theoretical analysis in §3 shows that a cheap-talk game in which the sender is insatiable and the sender and receiver strategy spaces are large is well-suited to compare the two models. In particular, such a context offers maximum scope for the models to make characteristically distinct predictions, for these differences in predictions to be empirically distinguishable, and also obviates the need to separately consider the predictions of the cognitive hierarchy model. Thus, we suggest this context is well-suited to conduct further comparisons of the two models.

In closing, this study represents a concrete step towards understanding what behavioral drivers are important to capture and explain observed behaviors in a practically relevant cheap-talk context. Our results encourage us to call for further research to compare the level-k and trust-embedded models in different business interactions and contexts. For example, one could examine cheap talk in other supply-chain settings with different payoff structures for the sender relative to the receiver (e.g, Özer et al. 2011, 2014), or study situations where the informed party provides advice rather than information (e.g., Özer et al. 2018), as well as situations where communication is not cheap talk (e.g., Inderfurth et al. 2013; Scheele et al. 2018). Comparing the two models in other business contexts can further shed light on how to design effective processes to facilitate and coordinate decisions across a value chain.

References

Amaldoss, W., S. Jain. 2005. Conspicuous consumption and sophisticated thinking. *Mgmt. Sci.* **51**(10) 1449–1466.

Amaldoss, W., S. Jain. 2010. Reference groups and product line decisions: An experimental investigation of limited editions and product proliferation. *Mgmt. Sci.* **56**(4) 621–644.

Angelova, V., T. Regner. 2013. Do voluntary payments to advisors improve the quality of financial advice? An experimental deception game. *J. Econ. Beh. Org.* **93**(September) 205–218.

Ashraf, N., I. Bohnet, N. Piankov. 2006. Decomposing trust and trustworthiness. *Exp. Econ.* **9**(3) 193–208.

Beer, R., H. Ahn, S. Leider. 2018. Can trustworthiness in a supply chain be signaled? *Mgmt. Sci.* **64**(9) 3974–3994.

Blume, A., D. V. DeJong, Y. Kim, G. B. Sprinkle. 2001. Evolution of communication with partial common interest. *Games Econ. Beh.* **37**(1) 79–120.

Bolton, G., B. Greiner, A. Ockenfels. 2013. Engineering trust: Reciprocity in the production of reputation information. *Mgmt. Sci.* **59**(2) 265–285.

Bostian, A. A., C. A. Holt, A. M. Smith. 2008. Newsvendor "pull-to-center" effect: Adaptive learning in a laboratory experiment. *Manuf. Serv. Oper. Mgmt.* **10**(4) 590–608.

Brehm, J., W. Rahn. 1997. Individual-level evidence for the causes and consequences of social capital. *Amer. J. Pol. Science* 999–1023.

Brinkhoff, A., Ö. Özer, G. Sargut. 2015. All you need is trust? An examination of inter-organizational supply chain projects. *Prodn. Oper. Mgmt.* **24**(2) 181–200.

Cai, H., J. T. Wang. 2006. Overcommunication in strategic information transmission games. *Games Econ. Beh.* **56**(1) 7–36.

Cain, D. M., G. Loewenstein, D. A. Moore. 2010. When sunlight fails to disinfect: Understanding the perverse effects of disclosing conflicts of interest. *J. Consumer Res.* **37**(5) 836–857.

Camerer, C. F., T. H. Ho, J. Chong. 2004. A cognitive hierarchy model of games. *Quart. J. Econ.* **119**(3) 861–898.

Campbell, M. C., A. Kirmani. 2000. Consumers' use of persuasion knowledge: The effects of accessibility and cognitive capacity on perceptions of an influence agent. *J. Consumer Res.* **27**(1) 69–83.

Cappelletti, D., W. Güth, M. Ploner. 2011. Being of two minds: Ultimatum offers under cognitive constraints. *J. Econ. Persp.* **32**(6) 940–950.

Chakraborty, A., R. Harbaugh. 2014. Persuasive puffery. *Mkt. Sci.* **33**(3) 382–400.

Chen, Y., X. Su, X. Zhao. 2012. Modeling bounded rationality in capacity allocation games with the quantal response equilibrium. *Mgmt. Sci.* **58**(10) 1952–1962.

Chu, L. Y., N. Shamir, H. Shin. 2016. Strategic communication for capacity alignment with pricing in a supply chain. *Mgmt. Sci.* **63**(12) 4366–4388.

Cornelissen, G., S. Dewitte, L. Warlop. 2011. Are social value orientations expressed automatically? Decision making in the dictator game. *Pers. Social Psych. Bulletin* **37**(8) 1080–1090.

Costa-Gomes, M. A., V. P. Crawford. 2006. Cognition and behavior in two-person guessing games: An experimental study. *Amer. Econ. Rev.* **96**(5) 1737–1768.

Crawford, V. P. 2003. Lying for strategic advantage: Rational and boundedly rational misrepresentation of intentions. *Amer. Econ. Rev.* **93**(1) 133–149.

Crawford, V. P., M. A. Costa-Gomes, N. Iriberri. 2013. Structural models of nonequilibrium strategic thinking: Theory, evidence, and applications. *J. Econom. Lit.* **51**(1) 5–62.

Crawford, V. P., N. Iriberri. 2007. Level-k auctions: Can a nonequilibrium model of strategic thinking explain the winner's curse and overbidding in private-value auctions? *Econometrica* **75**(6) 1721–1770.

Crawford, V. P., J. Sobel. 1982. Strategic information transmission. *Econometrica* **50**(6) 1431–1451.

Cui, T. H., P. Mallucci. 2016. Fairness ideals in distribution channels. *J. Marketing Res.* **53**(6) 969–987.

Cui, T. H., Y. Zhang. 2018. Cognitive hierarchy in capacity allocation games. *Mgmt. Sci.* **64**(3) 1250–1270.

Davis, A. M., E. Katok, N. Santamaría. 2014. Push, pull, or both? a behavioral study of how the allocation of inventory risk affects channel efficiency. *Mgmt. Sci.* **60**(11) 2666–2683.

Dickhaut, J., K. McCabe, A. Mukherji. 1995. An experimental study of strategic information transmission. *Econom. Theory* **6**(3) 389–403.

Doney, P. M., J. P. Cannon. 1997. An examination of the nature of trust in buyer-seller relationships. *J. Marketing* 35–51.

Donohue, K., Ö. Özer, Y. Zheng. 2020. Behavioral operations: Past, present, and future. *Manuf. Serv. Oper. Mgmt.* **22**(1) 191–202.

Ellingsen, T., R. Östling. 2010. When does communication improve coordination? *Amer. Econ. Rev.* **100**(4) 1695–1724.

Erat, S., U. Gneezy. 2012. White lies. *Mgmt. Sci.* **58**(4) 723–733.

Feng, T., Y. Zhang. 2017. Modeling strategic behavior in the competitive newsvendor problem: An experimental investigation. *Prodn. Oper. Mgmt.* **26**(7) 1383–1398.

Forsythe, R., R. Lundholm, T. Rietz. 1999. Cheap talk, fraud, and adverse selection in financial markets: Some experimental evidence. *Rev. Fin. Studies* **12**(3) 481–518.

Gardete, P. M. 2013. Cheap-talk advertising and misrepresentation in vertically differentiated markets. *Mkt. Sci.* **32**(4) 609–621.

Georganas, S., P. J. Healy, R. A. Weber. 2015. On the persistence of strategic sophistication. *J. Econ. Theory* **159** 369–400.

Gibson, R., C. Tanner, A. F. Wagner. 2013. Preferences for truthfulness: Heterogeneity among and within individuals. *Amer. Econ. Rev.* **103**(1) 532–548.

Gneezy, U. 2005. Deception: The role of consequences. *Amer. Econ. Rev.* **95**(1) 384–394.

Gneezy, U., A. Kajackaite, J. Sobel. 2018. Lying aversion and the size of the lie. *Amer. Econ. Rev.* **108**(2) 419–453.

Goldfarb, A., M. Xiao. 2011. Who thinks about the competition? managerial ability and strategic entry in us local telephone markets. *Amer. Econ. Rev.* **101**(7) 3130–61.

Goldfarb, A., B. Yang. 2009. Are all managers created equal? *J. Marketing Res.* **46**(5) 612–622.

Guo, T., S. Sriram, P. Manchanda. 2017. 'let the sun shine in': The impact of industry payment disclosure on physician prescription behavior. Working Paper, University of Michigan.

Hardin, R. 2002. *Trust and Trustworthiness*. Russell Sage Foundation.

Haruvy, E., D. O. Stahl, P. W. Wilson. 2001. Modeling and testing for heterogeneity in observed strategic behavior. *Rev. Econom. & Stat.* **83**(1) 146–157.

Hastie, T., R. Tibshirani, J. Friedman. 2011. *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. New York: Springer.

Ho, T. H., N. Lim, T. H. Cui. 2010. Reference dependence in multilocation newsvendor models: A structural analysis. *Mgmt. Sci.* **56**(11) 1891–1910.

Ho, T. H., X. Su. 2013. A dynamic level-k model in sequential games. *Mgmt. Sci.* **59**(2) 452–469.

Ho, T. H., J. Zhang. 2008. Designing pricing contracts for boundedly rational customers: Does the framing of the fixed fee matter? *Mgmt. Sci.* **54**(4) 686–700.

Hossain, T., J. Morgan. 2013. When do markets tip? a cognitive hierarchy approach. *Mkt. Sci.* **32**(3) 431–453.

Inderfurth, K., A. Sadrieh, G. Voigt. 2013. The impact of information sharing on supply chain performance under asymmetric information. *Prodn. Oper. Mgmt.* **22**(2) 410–425.

Jin, G. Z., M. Luca, D. Martin. 2018. Is no news (perceived as) bad news? an experimental investigation of information disclosure. NBER Working Paper.

Kahneman, D., A. Tversky. 1982. On the study of statistical intuitions. *Cognition* **11**(2) 123–141.

Katok, E., V. Pavlov. 2013. Fairness in supply chain contracts: A laboratory study. *J. Ops. Mgmt.* **31**(3) 129–137.

Katok, E., D. Y. Wu. 2009. Contracting in supply chains: A laboratory investigation. *Mgmt. Sci.* **55**(12) 1953–1968.

Kawagoe, T., H. Takizawa. 2009. Equilibrium refinement vs. level-k analysis: An experimental study of cheap-talk games with private information. *Games Econ. Beh.* **66**(1) 238–255.

Kawagoe, T., H. Takizawa. 2012. Level-k analysis of experimental centipede games. *J. Econ. Beh. Org.* **82**(2-3) 548–566.

Kessler, J. B., S. Leider. 2012. Norms and contracting. *Mgmt. Sci.* **58**(1) 62–77.

Li, J., D. R. Beil, S. Leider. 2019. Team decision making in operations management Working Paper, University of Michigan.

Lim, N., H. Chen. 2014. When do group incentives for salespeople work? *J. Marketing Res.* **51**(3) 320–334.

Lim, N., S. H. Ham. 2014. Relationship organization and price delegation: An experimental study. *Mgmt. Sci.* **60**(3) 586–605.

Lim, N., T. H. Ho. 2007. Designing price contracts for boundedly rational customers: Does the number of blocks matter? *Mkt. Sci.* **26**(3) 312–326.

Loch, C. H., Y. Wu. 2008. Social preferences and supply chain performance: An experimental study. *Mgmt. Sci.* **54**(11) 1835–1849.

Lundquist, T., T. Ellingsen, E. Gribbe, M. Johannesson. 2009. The aversion to lying. *J. Econ. Beh. Org.* **70**(1-2) 81–92.

Mazar, N., O. Amir, D. Ariely. 2008. The dishonesty of honest people: A theory of self-concept maintenance. *J. Marketing Res.* **45**(6) 633–644.

McKelvey, R. D., T. R. Palfrey. 1998. Quantal response equilibria for extensive form games. *Exp. Econ.* **1**(1) 9–41.

Michaely, R., K. L. Womack. 1999. Conflict of interest and the credibility of underwriter analyst recommendations. *Rev. Fin. Studies* **12**(4) 653–686.

Moorman, C., R. Deshpande, G. Zaltman. 1993. Factors affecting trust in market research relationships. *J. Marketing* 81–101.

Nagel, R. 1995. Unraveling in guessing games: An experimental study. *Amer. Econ. Rev.* **85**(5) 1313–1326.

Özer, Ö., U. Subramanian, Y. Wang. 2018. Information sharing, advice provision, or delegation: What leads to higher trust and trustworthiness? *Mgmt. Sci.* **64**(1) 474–493.

Özer, Ö., Y. Zheng. 2019. *The Handbook of Behavioral Operations*, chap. Trust and Trustworthiness. Wiley Series in Operations Research and Management Science, 489–524.

Özer, Ö., Y. Zheng, K. Chen. 2011. Trust in forecast information sharing. *Mgmt. Sci.* **57**(6) 1111–1137.

Özer, Ö., Y. Zheng, Y. Ren. 2014. Trust, trustworthiness, and information sharing in supply chains bridging China and the United States. *Mgmt. Sci.* **60**(10) 2435–2460.

Rand, D. G., J. D. Greene, M. A. Nowak. 2012. Spontaneous giving and calculated greed. *Nature* **489**(7416) 427–430.

Sánchez-Pagés, S., M. Vorsatz. 2007. An experimental study of truth-telling in a sender-receiver game. *Games Econ. Beh.* **61**(1) 86–112.

Sánchez-Pagés, S., M. Vorsatz. 2009. Enjoy the silence: An experiment on truth-telling. *Exp. Econ.* **12**(2) 220–241.

Scheele, L. M., U. W. Thonemann, M. Slikker. 2018. Designing incentive systems for truthful forecast information sharing within a firm. *Mgmt. Sci.* **64**(8) 1–24.

Schulz, J. F., U. Fischbacher, C. Thöni, V. Utikal. 2014. Affect and fairness: Dictator games under cognitive load. *J. Econ. Persp.* **41** 77–87.

Schwartz, J., M. F. Luce, D. Ariely. 2011. Are consumers too trusting? the effects of relationships with expert advisers. *J. Marketing Res.* **48**(SPL) S163–S174.

Sheremeta, R. M., T. W. Shields. 2013. Do liars believe? beliefs and other-regarding preferences in sender-receiver games. *J. Econ. Beh. Org.* **94** 268–277.

Sobel, J. 2013. Ten possible experiments on communication and deception. *J. Econ. Beh. Org.* **93** 408–413.

Spiliotopoulou, E., K. Donohue, M. c. Gürbüz. 2016. Information reliability in supply chains: The case of multiple retailers. *Prodn. Oper. Mgmt.* **25**(3) 548–567.

Stahl, D. O., P. W. Wilson. 1994. Experimental evidence on players' models of other players. *J. Econ. Beh. Org.* **25**(3) 309–327.

Stahl, D. O., P. W. Wilson. 1995. On players' models of other players: Theory and experimental evidence. *Games Econ. Beh.* **10**(1) 218–254.

Terwiesch, C., Z. J. Ren, T. H. Ho, M. A. Cohen. 2005. An empirical analysis of forecast sharing in the semiconductor equipment supply chain. *Mgmt. Sci.* **51**(2) 208–220.

Wang, J. T., M. Spezio, C. F. Camerer. 2010. Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *Amer. Econ. Rev.* **100**(3) 984–1007.

Zaki, J., J. P. Mitchell. 2013. Intuitive prosociality. *Current Directions in Psych Sci.* **22**(6) 466–470.

Zhou, B., C. F. Mela, W. Amaldoss. 2015. Do firms endowed with greater strategic capability earn higher profits? *J. Marketing Res.* **52**(3) 325–336.

Appendix A Proofs for Theorems

Proof for Theorem 2: Maximizing the receiver's expected payoff in Equation (5), we obtain $a^*\left(\hat{\xi};\alpha_R\right)$ as stated in the theorem. For the sender, given her belief about the receiver's trust type, we have

$$\mathbf{E}\left[a^*\left(\hat{\xi};\alpha_R\right) \mid H\left(\alpha_S\right)\right] = \frac{r}{c}\left[\frac{\left(1+\bar{\alpha}_S\right)\hat{\xi} + \left(1-\bar{\alpha}_S\right)\underline{\xi}}{2}\right], \tag{16}$$

where $\bar{\alpha}_S$ is defined in the statement of the theorem. Therefore, from Equation (7), the sender's expected payoff is

$$\Pi_S\left(\hat{\xi},\xi;\gamma,H\left(\alpha_s\right)\right)=s\xi\left(\frac{r}{c}\left[\frac{\left(1+\bar{\alpha}_S\right)\hat{\xi}+\left(1-\bar{\alpha}_S\right)\underline{\xi}}{2}\right]\right)-G\left(\left|\hat{\xi}-\xi\right|;\gamma\right), \tag{17}$$

Note that $\Pi_S\left(\hat{\xi},\xi;\gamma,H\left(\alpha_s\right)\right)$ is strictly increasing in $\hat{\xi}$ for $\hat{\xi}\leq\xi$, and strictly concave in $\hat{\xi}$ for $\hat{\xi}>\xi$. Hence, there is a unique optimal message $\hat{\xi}^*\left(\xi;\gamma,H\left(\alpha_s\right)\right)\geq\xi$. From Equation (17), for $\hat{\xi}>\xi$,

$$\frac{\partial\Pi_S\left(\hat{\xi},\xi;\gamma,H\left(\alpha_s\right)\right)}{\partial\hat{\xi}}=s\frac{r}{c}\left(\tfrac{1+\bar{\alpha}_S}{2}\right)\xi-g\left(\hat{\xi}-\xi;\gamma\right). \tag{18}$$

Since $g\left(0;\gamma\right)=0$, it follows that $\hat{\xi}^*\left(\xi;\gamma,H\left(\alpha_s\right)\right)=\min\left\{\bar{\xi},A\left(\gamma,\bar{\alpha}_S\right)\cdot\xi\right\}$ where $A\left(\gamma,\bar{\alpha}_S\right)$ is defined in the statement of the theorem.

Proof for Theorem 3: We prove this result inductively. An $L0$ receiver is fully-believing and hence will take the fully-believing action, i.e., $a_0\left(\hat{\xi}\right)=a_I\left(\hat{\xi}\right)$. Suppose an $L\left(k-1\right)$ receiver believes messages $\hat{\xi}\leq\bar{\xi}-k+1$ and takes action $a_I\left(\hat{\xi}\right)$, and disregards all higher messages $\hat{\xi}>\bar{\xi}-k+1$ and takes action a_{NI}. Now, an Lk sender's expected payoff is

$$\Pi_{Sk}\left(\hat{\xi},\xi\right)=\begin{cases}s\xi a_I\left(\hat{\xi}\right)=s\xi\frac{r}{c}\hat{\xi}, & \hat{\xi}\leq\bar{\xi}-k+1;\\ s\xi a_{NI}=s\xi\frac{r}{c}\frac{\left(\bar{\xi}+\underline{\xi}\right)}{2}, & \hat{\xi}>\bar{\xi}-k+1.\end{cases} \tag{19}$$

Consequently, regardless of her true information ξ, the Lk sender sends the highest message that the $L(k-1)$ receiver will believe, namely $\hat{\xi}_k\left(\xi\right)=\bar{\xi}-\left(k-1\right)$, provided doing so would induce a higher action than a_{NI}; if all messages that the receiver believes will induce an action less than or equal to a_{NI}, then the Lk sender sends the lowest message that induces a_{NI}, namely $\hat{\xi}_k\left(\xi\right)=\frac{\bar{\xi}+\underline{\xi}}{2}$. Therefore, $\hat{\xi}_k\left(\xi\right)=\hat{\xi}_k=\max\left\{\bar{\xi}-\left(k-1\right),\frac{\bar{\xi}+\underline{\xi}}{2}\right\}$. An Lk receiver will believe all messages $\xi\leq\tilde{\xi}_k=\hat{\xi}_k-1$, taking them to be from an $L0$ sender; and will disregard all higher messages $\xi>\tilde{\xi}_k$, taking them to be from the appropriate sender type from $L1$ to Lk, and hence uninformative. Therefore the Lk receiver's expected payoff is

$$\Pi_{Rk}\left(a,\hat{\xi}\right)=\begin{cases}r\hat{\xi}a-\frac{1}{2}ca^2, & \hat{\xi}\leq\tilde{\xi}_k;\\ r\frac{\left(\bar{\xi}+\underline{\xi}\right)}{2}a-\frac{1}{2}ca^2, & \hat{\xi}>\tilde{\xi}_k\end{cases} \tag{20}$$

Hence, the Lk receiver takes action $a_I\left(\hat{\xi}\right)$ for $\hat{\xi}\in\left[\underline{\xi},\tilde{\xi}_k\right]$, and action a_{NI} otherwise.

Appendix B Experimental Setup in Prior Cheap Talk Experiments

The table below describes whether in previous cheap-talk experiments the sender is insatiable and the size of the sender and receiver strategy space. We exclude deception games in which receiver does

not know sender's payoff (e.g., Gneezy 2005; Mazar et al. 2008), capacity allocation games in which there are multiple senders (Spiliotopoulou et al. 2016) or the receiver's strategy is exogeneously fixed (Cui and Zhang 2018), or games in which the receiver chooses the distribution of his posterior belief (not action) in response to the sender's message (e.g., Inderfurth et al. 2013).

	Insatiable Sender	# of Sender Messages	# of Receiver Actions
Dickhaut et al. (1995)	No	4	4
Forsythe et al. (1999)	Yes	3	3
Blume et al. (2001)	No	3	5
Cai and Wang (2006)	No	3	3
Sánchez-Pagés and Vorsatz (2007)	Yes	2	2
Kawagoe and Takizawa (2009)	Yes	2	3
Lundquist et al. (2009)	No	100	2
Sánchez-Pagés and Vorsatz (2009)	Yes	2	2
Wang et al. (2010)	No	5	5
Özer et al. (2011)	No	301	301
Sheremeta and Shields (2013)	Yes	2	2
Özer et al. (2014)	No	301	301
Özer et al. (2018)	Yes	71	201

We observe that most past experiments use fairly small strategy spaces with a handful of messages and actions. Thus, they do not provide sufficient scope to reliably distinguish between the predictions of the two models. Essentially, such experiments were designed to distinguish between the standard theory prediction and a behavioral economics theory, but are not well-suited for comparing two behavioral models, in particular, the level-k and trust-embedded models. We remark that Özer et al. (2011) and Özer et al. (2014) use experiments with relatively large strategy spaces for senders and receivers. However, the sender's payoff in their context is *not* strictly increasing for the entire range of receiver's possible actions. As a result, under the level-k model, the sender's pecuniary payoff structure can also constrain the extent to which the sender distorts the message. We conjecture that distinguishing the two behavioral models in this case would be more difficult relative to the cheap-talk context in Özer et al. (2018), which the present paper uses. Another experimental paper that has large strategy spaces is Scheele et al. (2018). The authors mainly focus on experiments with costly *non*-cheap-talk communication. Nevertheless, their experiment design

includes a benchmark treatment in which communication is cheap talk and the sender is insatiable. However, this single treatment provides a small dataset: 8 rounds of interaction between 16 participants. In contrast, Özer et al. (2018) provide 11 rounds of interaction between 60 participants, and also includes an experimental manipulation.

Appendix C Trust-Embedded Model: Estimable Sender Parameters

The sender's expected payoff from sending message $\hat{\xi}$ given actual information ξ is

$$\Pi_S\left(\hat{\xi},\xi;\gamma,H\left(\alpha_S\right)\right)=s\xi\mathbf{E}\left[a^*\left(\hat{\xi};\alpha_R\right)\mid H\left(\alpha_S\right)\right]-\frac{1}{2}\gamma\left(\hat{\xi}-\xi\right)^2, \tag{21}$$

where

$$\mathbf{E}\left[a^*\left(\hat{\xi};\alpha_R\right)\mid H\left(\alpha_S\right)\right]=\frac{r}{c}\left[\frac{\left(1+\bar{\alpha}_S\right)\hat{\xi}+\left(1-\bar{\alpha}_S\right)\underline{\xi}}{2}\right]. \tag{22}$$

As described in §4, a type ω sender's logit utility from sending message $\hat{\xi}$ given actual information ξ is the sum of her theoretical predicted payoff and the logit shock ϵ, given by

$$\begin{aligned}\Pi_S\left(\hat{\xi},\xi;\gamma_\omega,\bar{\alpha}_{S\omega},\lambda_\omega\right)+\frac{1}{\lambda_\omega}\epsilon&=s\xi\left(\frac{r}{c}\left[\frac{(1+\bar{\alpha}_{S\omega})\hat{\xi}+(1-\bar{\alpha}_{S\omega})\underline{\xi}}{2}\right]\right)-\frac{1}{2}\gamma_\omega\left(\hat{\xi}-\xi\right)^2+\frac{1}{\lambda_\omega}\epsilon.\\&=C\left(\xi\right)+F\left(\hat{\xi},\xi;A\left(\gamma_\omega,\bar{\alpha}_{s\omega}\right)\right)\gamma_\omega+\frac{1}{\lambda_\omega}\epsilon,\end{aligned} \tag{23}$$

where $C\left(\xi\right)=s\frac{r}{c}\frac{(1-\bar{\alpha}_{S\omega})}{2}\underline{\xi}\xi-\frac{1}{2}\gamma_\omega\xi^2$ is independent of $\hat{\xi}$, and $F\left(\hat{\xi},\xi;A\left(\gamma_\omega,\bar{\alpha}_{s\omega}\right)\right)=A\left(\gamma_\omega,\bar{\alpha}_{s\omega}\right)\cdot\xi\hat{\xi}-\frac{1}{2}\hat{\xi}^2$. Substituting from Equation (23) in Equation (12), the logit likelihood of message $\hat{\xi}_{it}$ is

$$Pr_S(\hat{\xi}_{it}\mid\xi_t,\gamma_\omega,\bar{\alpha}_{s\omega},\lambda_\omega)=\frac{\exp\{\lambda_\omega\gamma_\omega F\left(\hat{\xi}_{it},\xi_t;A\left(\gamma_\omega,\bar{\alpha}_{s\omega}\right)\right)\}}{\sum_{\hat{\xi}\in\Delta_S}\exp\{\lambda_\omega\gamma_\omega F\left(\hat{\xi},\xi_t;A\left(\gamma_\omega,\bar{\alpha}_{s\omega}\right)\right)\}}, \tag{24}$$

Hence, only $A\left(\gamma_\omega,\bar{\alpha}_{S\omega}\right)$ and $\lambda_\omega\gamma_\omega$ affect the logit choice probabilities.

Appendix D Predicted vs. Observed Behaviors

Figure 1: Predicted vs. Observed Sender Behaviors

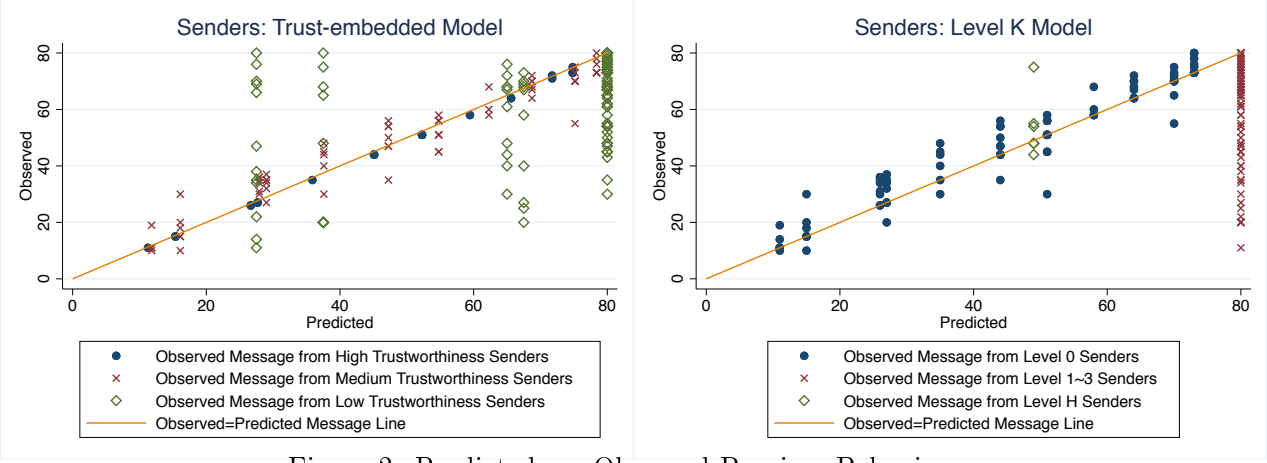

Figure 2: Predicted vs. Observed Receiver Behaviors