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

Xiaolin Li, Özalp Özer, Upender Subramanian
The University of Texas at Dallas, Richardson (TX), {xiaolin.li, o.ozer, 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 level-k model and the trust-embedded model, have emerged to explain how cheap talk works between two decision makers. The level-k model considers decision makers to be boundedly rational in their strategic thinking. In contrast, the trust-embedded model considers decision makers to be motivated by non-pecuniary motives to be trusting and trustworthy. While both theories have been successful in explaining cheap talk in separate contexts, they point to contrasting underlying drivers for human behavior. In this paper, we provide the first direct comparison of both theories within the same unified context. We show that in a cheap-talk context that well represents many practical business and economic situations, the level-k and trust-embedded models make characteristically distinct and empirically distinguishable predictions. We leverage past experimental data from this context to structurally estimate both models, and let the data inform us about which model has better explanatory power and predictive performance for observed behaviors. Our findings shed light on the behavioral drivers of cheap talk, and can inform academics and practitioners in designing systems and processes to improve the outcomes of cheap-talk interactions in business and economic settings.*

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

History: This version January 2019.

1 Introduction

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

* The authors thank Ernan Haruvy, Karen Zheng, and participants in their presentation at the 2018 Behavioral Operations Management Conference for their constructive comments. The authors’ names are listed in alphabetical order.
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., Michael 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 many other practical situations, the information communicated by the better-informed party can be easy to distort, difficult to verify, and does not per se create binding commitments on the communicating party. Consequently, “talk is cheap”, and the communication need not be truthful.

In fact, there is often a monetary incentive for cheap talk to be deceptive, which raises questions about the value and effectiveness of cheap talk between parties with conflicting interests. For example, advertisers earn revenues and salespeople earn commissions from convincing consumers to buy; downstream firms realize higher sales and profit if there is abundant upstream supply; financial managers make commissions if investors invest; and physicians bill for procedures performed and may receive perks from pharmaceutical companies. Therefore, the interests of the parties involved are never fully aligned. Nevertheless, cheap talk is widely used between parties with conflicting interests. Furthermore, 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., Michael 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 the underlying drivers of cheap-talk interactions and outcomes.

Experimental research has consistently found that standard economic theory does not provide an adequate account of cheap-talk interactions between human decision makers (e.g., Cai and Wang 2006; Özer et al. 2011; Erat and Gneezy 2012; Spiliotopoulos et al. 2016). 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 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). The other theory is based on a more positive view that individuals are not only guided by their self-interest, but also by non-pecuniary motives to be trusting and trustworthy in their interactions with others (e.g., Gneezy 2005; Mazar et al. 2008; Özer et al. 2011; Schwartz et al. 2011). While both theories are reasonable in many practical situations, and have been successful in explaining behaviors in separate contexts, they point to contrasting underlying drivers for cheap-talk interactions and outcomes.
In this paper, we conduct the first direct comparison of both behavioral theories within the same unified context to shed light on which of them provides a better account of observed behaviors. 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 to develop processes and establish more effective cheap talk interactions that lead to profitable outcomes.

**Standard Economic Model of Cheap Talk and Experimental Findings.** Standard economic analysis predicts that, while cheap talk between parties with conflicting interests will not be truthful, it can nevertheless coordinate decisions to some extent under certain conditions. The canonical framework for studying cheap talk introduced by Crawford and Sobel (1982) consists of a sender (she) who sends a cheap-talk message to a receiver (he) prior to the receiver taking an action that affects both their payoffs. The sender's message can reveal her private information about an underlying state of the world that is relevant for the receiver's action. The sender's and receiver's interests are, however, conflicting. Specifically, Crawford and Sobel (1982) examine a setting in which, for any underlying state of the world, the receiver action that is optimal for the sender differs from what is optimal for the receiver by a known fixed amount. The size of this fixed amount captures the degree of the conflict in their interests. Consequently, the sender has an incentive not to be truthful, and to distort her message in an attempt to influence the receiver to take an action that benefits the sender at the receiver's expense. Hence, the sender's message is cheap talk. Formally, cheap talk is defined as communication that is costless, non-verifiable and non-binding. In particular, it is either too costly or impractical to write and enforce a contract to ensure truthful communication. Cheap talk is said to be informative if it reduces the uncertainty about the uncertain underlying state, and influential if it affects the receiver's actions.

Crawford and Sobel (1982) show that cheap talk can be partially informative and influential in this setting if the sender's and receiver's interests are not too conflicting, i.e., the difference between the sender's and receiver's preferred actions (for any underlying state of the world) is not too large. Intuitively, if it is in the sender's interest not to mislead the receiver by much, then the message can still be informative, albeit distorted; and, it can be worthwhile for the receiver to respond to the sender's message taking the distortion into account. The extent to which the sender's message is informative and influential is determined in equilibrium: standard economic analysis assumes that the receiver correctly anticipates the distortion in a message and accounts for it as per standard
Bayesian inference; and, the sender correctly anticipates how the receiver responds to distorted messages, and decides how much to distort a message. A further theoretical prediction is that the communication is more informative and influential if the sender’s and receiver’s interests are less conflicting.

Experimental research has confirmed that cheap talk between parties with conflicting interests can be (partially) informative and influential, and even more so if the sender and receiver interests are less conflicting (e.g., Blume et al. 2001; Cai and Wang 2006). However, a consistent finding is that there is substantial and systematic over-communication: sender messages are more informative, and receivers are more influenced by the messages than predicted by standard theory (e.g., Cai and Wang 2006; Mazar et al. 2008; 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 some senders being more truthful than others, 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

**Behavioral Economics Models of Cheap Talk.** Two separate streams of behavioral economics research have developed to explain the observed over-communication phenomenon. The first stream of research considers decision makers to be limited in their ability to think strategically, and has modeled this behavior using a level-k model (Crawford 2003; Cai and Wang 2006; Kawagoe and Takizawa 2009; Ellingsen and Östling 2010; Wang et al. 2010).² The level-k model was originally proposed by Nagel (1995) and Stahl and Wilson (1994, 1995) as a non-equilibrium model of strategic behavior. It has been used to explain deviations from standard theory predictions in a wide variety of games (see Crawford et al. 2013 and Georganas et al. 2015 for recent surveys of level-k model applications). In this model, players differ in their type or level of strategic thinking. A higher-level player type correctly anticipates and accounts for the behavior of lower-level player types and is thus more sophisticated in his or her strategic thinking than a lower-level player type. The lowest level of strategic thinking is denoted by $L0$, and progressively higher levels are denoted by $L1$, $L2$, $L3$ and so on. Each player type believes that other players are less sophisticated (are of lower types) than him or herself and best responds to this belief. In particular, many applications of the level-k

¹We 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.

model (including cheap-talk games) consider that each higher type best responds to the belief that other players are of the immediately lower type (e.g., see review by Crawford et al. 2013). Therefore, $L_1$ best responds to $L_0$, $L_2$ best responds to $L_1$ and so on. A higher-level type anticipates the behavior of lower types through iterative thinking about the strategies of successive levels of player types starting from the $L_0$ type. The $L_0$ type thus serves as the starting point for each higher type’s model of others: the $L_0$ type is taken to be non-strategic, following a behavior that represents naive or unsophisticated play. Level-k models of cheap talk assume that an $L_0$ sender is truthful and does not distort the message, and an $L_0$ receiver is credulous and believes all messages to be truthful (Cai and Wang 2006; Kawagoe and Takizawa 2009; Wang et al. 2010).

The level-k model is able to explain the observed behaviors, given the truthful and credulous behaviors of the $L_0$ type, and the behaviors of successive higher types, which are anchored on the behaviors of the $L_0$ type. In particular, each higher-level sender distorts the message to influence a lower-level receiver, who is more influenced by sender messages than the fully-strategic receiver under standard theory; essentially, a boundedly-strategic receiver in the level-k model does not fully anticipate the extent to which the sender, especially of a higher type than the receiver, may have distorted the message. Consequently, the level-k model is able to explain the over-communication phenomenon. Furthermore, the observed heterogeneity in sender and receiver behaviors can be explained by players’ levels of thinking: different sender types distort the message to different extents, and different receiver types are influenced by messages to different extents. Cai and Wang (2006) find that the observed behaviors in their experiments could be explained by the level-k model as well as by 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. 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 most attends to is consistent with

---

3 Later, in §3.2, we discuss the cognitive hierarchy model proposed by Camerer et al. (2004) in which each higher-level type believes that the other player is drawn from a (truncated Poisson) distribution of lower level types.

4 Crawford et al. (2013) (pg. 51) explain the assumption about $L_0$ type as follows: “In games with communication, it would be behaviorally odd if a [receiver’s] strategically naive assessment of a message did not initially favor its literal interpretation, even if he [subsequently] ends up not taking it at face value.” The $L_0$ receiver’s credulous behavior can also be seen as a best response to the $L_0$ sender’s truthful behavior. In other (non cheap-talk) games, the $L_0$ type is often taken to be uniformly randomizing across all actions. However, assuming that the $L_0$ sender uniformly randomizes across all messages does not explain observed behaviors in cheap-talk games.
her estimated level of thinking.

The second stream of research explains cheap-talk behavior based on individuals’ social preferences to be trusting and trustworthy. Accumulating real-world and laboratory evidence has shown that decision makers are more truthful than predicted by standard theory (at the expense of their monetary gain) 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). Essentially, decision makers appear to tradeoff their monetary gain from lying against a non-pecuniary cost from lying. Researchers have further found that decision makers are more trusting of others than predicted by standard theory in a variety of situations (e.g., Forsythe et al. 1999; Cain et al. 2010; Özer et al. 2011; Schwartz et al. 2011). Essentially, decision makers appear to place a higher belief in others acting in a trustworthy manner than expected for a “calculative” decision maker under standard theory and Bayesian inference. Moreover, consistent with the “experiential view” of trust and trustworthiness, 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 2018).

The combination of trusting and trustworthy behaviors by senders and receivers respectively can also explain the over-communication phenomenon. Senders are more truthful than predicted by standard theory because they face a lying cost that discourages them from distorting their message for monetary gain. Receivers are more influenced by sender messages than predicted by standard theory because they have a propensity to trust, effectively following a simpler rule than implied by Bayesian inference to update their beliefs about the sender’s private information based on her message. Furthermore, since propensity to trust or to be trustworthy varies across individuals (e.g., Ashraf et al. 2006; Gibson et al. 2013; Özer et al. 2018), there is heterogeneity in individuals behaviors. Sánchez-Pagès and Vorsatz (2007) find that cost of lying can explain sender’s behavior better than the AQRE model. Özer et al. (2011) propose and test a trust-embedded model of cheap talk that incorporates cost of lying for senders and a trust-based belief-updating rule for receivers. Their trust-embedded model 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.
Our Study. In this paper, we provide the first direct comparison of the two behavioral economics theories of cheap talk in a unified context. Both theories have been successful in explaining cheap-talk behaviors in separate contexts. Both 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 of limited strategic thinking they underestimate the extent to which the salesperson’s strategic intent matters 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 how strategic the consumer is in anticipating the strategic intent of the salesperson. The two theories, however, point to contrasting underlying drivers for human behavior 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 and/or social uncertainties surrounding a sender are likely to be more effective in improving cheap-talk interactions (Özer and Zheng 2018). Therefore, comparing the two behavioral theories of cheap talk within the same unified context is of theoretical and practical interest. This comparison also brings us one step closer to building a unified theory that explains the (effective or ineffective) role of cheap talk in facilitating business and economic transactions between parties with conflicting interests.

Our objective, therefore, is to determine which behavioral economics theory on its own provides a better account of observed behaviors in the same cheap-talk context and corresponding experiments. We leverage existing cheap-talk experiment data from Özer et al. (2018) because the authors share this data with scholars (hence it is readily available) and, as we elaborate in §2, the context and experiments well represent many practical situations and are well-suited for comparing the two behavioral models. Specifically, we develop the level-k and trust-embedded model predictions for the cheap-talk game in Özer et al. (2018). We find that both models predict distorted yet informative and influential communication, and both models can potentially explain the over-communication phenomenon and heterogeneity in sender and receiver behaviors. Despite these similarities, however, the two models lead to different implications for how people arrive at their decisions (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, the two models predict distinct patterns of individual-level behaviors, which make it possible to empirically distinguish between them based on
how well they account for observed behaviors. We then let the data inform us as to which model explains observed behaviors better: we structurally estimate both models with the experiment data, and compare their performance with respect to in-sample fit, out-of-sample predictions, and ability to recover the effect of an experimental manipulation.\(^5\)

Overall, our results offer support for the more positive view of human behavior, namely, that individuals are not only guided by their pecuniary self-interest, but also by their non-pecuniary motives of trust and trustworthiness. The trust-embedded model performs better than the level-k model in explaining observed behaviors of senders as well as receivers. Specifically, the tendency for senders messages to be distorted yet informative is better explained by senders’ psychological lying costs under the trust-embedded model than iterative levels of strategic thinking under the level-k model. Similarly, the tendency for receivers to be influenced by sender messages is better explained by a trust-based inference process (a non-Bayesian inference process) under the trust-embedded model that captures the receiver’s innate tendency to trust, than a Bayesian inference process based on iterative levels of strategic thinking under the level-k model. Our results thus highlight the importance of designing processes to reduce barriers for and engineer trusting and trustworthy relationships for cheap-talk interactions.

Our work also 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. Specifically, 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.\(^6\) Another literature stream has examined the role of trust and trustworthiness due to non-pecuniary motives.

---

\(^5\) 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 Iribarri (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.

\(^6\) 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 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).

In what follows, in §2, we describe the cheap-talk game in Özer et al. (2018). In §3, we formally 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 describe the experiment data, and our model estimation and comparison approach. In §5, we present our results and observations. In §6, we conclude with a discussion on the implications of our findings.

2 A Cheap-Talk Context to Compare Level-k and Trust Models
As noted in Sobel (2013), distinguishing between boundedly-strategic behavior and non-pecuniary preferences can be challenging in cheap-talk games as they both can lead to qualitatively similar predictions. We find that three features of the cheap-talk game in Özer et al. (2018) make it well-suited for comparing the two behavioral models. First, the sender and receiver strategy spaces are large: the number of possible sender messages is 81 and receiver actions is 181. Intuitively, if the strategy spaces are small, then there is limited room for the predictions of one model to differ from that of the other, which makes it difficult to reliably distinguish between these predictions empirically, especially after accounting for random errors in participant decisions. In contrast, a large strategy space makes it easier to observe the differences in predicted behaviors. Second, the sender’s pecuniary payoff is strictly increasing in the receiver’s action over the entire range of receiver ac-

7 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.
tions. Therefore, the sender strictly prefers that the receiver take as high an action regardless of the sender’s private information. In particular, under either model, the sender’s pecuniary payoff does not constrain the sender’s incentive to distort the message, which leads to different implications for how individuals arrive at their decisions under each model. As a result, the two models predict characteristically distinct individual-level behaviors that can be empirically distinguished. Finally, a further implication of the strictly monotone sender payoff is that the cognitive hierarchy model, which considers that each higher-level player believes that other players are drawn from a distribution of lower-level types, predicts similar patterns of behavior as the level-k model. As a result, it is not necessary to separately consider such alternative models of boundedly-strategic thinking. We formally derive these model predictions and elaborate rigorously on each of the aforementioned issues in §3 and throughout the paper. We further remark that, to our knowledge, only Özer et al. (2018) use cheap talk experiments that incorporate large strategy spaces and strictly monotone sender payoff; we compare the experimental setups of prior cheap talk experiments later in §6 and Appendix C.

The cheap-talk game in Özer et al. (2018) study the communication between a supplier (e.g., P&G) and a retailer (e.g., Kroger). In this context, the supplier is better informed about its product’s market (i.e., sales) potential than the retailer but is still uncertain about final sales. The retailer needs the supplier’s private information to optimally decide how much store resources, such as shelf-space, to allocate to the product. The supplier prefers the retailer to allocate the maximum amount of available store resources, irrespective of its product’s actual market potential, because higher store resources result in higher sales and profits for the supplier. However, store resources are costly for the retailer who, therefore, prefers to allocate more resources to a product with higher market potential. Knowing this fact, the supplier has an incentive to exaggerate its product’s market potential when communicating to the retailer.

More formally, the market potential for the supplier’s product is uncertain, given by $q = \xi + \epsilon$, where $\xi$ is a positive integer uniformly distributed over the support $[\xi, \bar{\xi}]$, and $\epsilon$ is an integer that is uniformly distributed over the support $[-e, e]$. The variable $\xi$ is known to the supplier but not to the retailer, and represents the supplier’s private demand information; the other variable $\epsilon$ 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 $\xi$ and sends a report $\hat{\xi} \in [\underline{\xi}, \bar{\xi}]$ 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 $\epsilon$ 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 $C(a) = \frac{1}{2}ca^2$. Thus, the supplier’s and retailer’s profits are respectively:

$$\Pi_S = sD(a) = sqa$$

$$\Pi_R = rD(a) - C(a) = rqa - \frac{1}{2}ca^2.$$  

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’s profit is strictly increasing in the retailer’s service level $a$, i.e., sender’s payoff is strictly monotone in the receiver’s action. Therefore, the supplier benefits from inducing as high a service decision as possible. Thus, the interests of the two parties are not aligned. Moreover, any communication regarding the product’s market potential is essentially a cheap talk (i.e., the supplier’s report of $\xi$ is costless, non-binding and not verifiable).

Since our interest in understanding cheap-talk behaviors extends beyond the above distribution channel setting, in the rest of the paper, we abstract away from this specific setting and instead use the terminology of a sender-receiver cheap-talk game; specifically, denoting the supplier as the sender, and the retailer as the receiver. We further remark that cheap-talk games in which the sender’s payoff is strictly increasing in the receiver’s action, as in the game above, 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. We refer to such cheap-talk games in which the sender’s payoff is strictly increasing in the receiver’s action as cheap talk by an insatiable sender, i.e., the sender is not “satisfied” no

---

8We deviate from 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.
matter how high the receiver’s action is.

3 Standard And Behavioral Economics Theory Predictions

In this section, we first present the standard economic model of the aforementioned cheap-talk game, and summarize what this theory predicts (as was originally discussed in Özer et al. 2018). Next, we develop the level-k and the trust-embedded models of this cheap-talk game. We characterize the resulting decisions and outcomes under each behavioral model and show what each model predicts an individual (sender or receiver) would do in this context.

3.1 Standard Economic Theory Predictions

As described in §2, the sender’s (e.g., supplier’s) and receiver’s (e.g., retailer’s) payoff are respectively given by,

\[ \Pi_S = sqa, \]
\[ \Pi_R = rqa - \frac{1}{2}ca^2, \]

where \( q = \xi + \epsilon \). Let \( \hat{\xi}(\xi) \) denote the sender’s message strategy after observing \( \xi \). We observe from Equation (1) that the sender would prefer the receiver to pick the highest possible level of action, i.e., \( a = \bar{a} \), irrespective of \( \xi \). Let \( a_{NI} \) denote the receiver’s uninformed optimal decision based only on the knowledge of the prior distribution of \( q \). Let \( a_I(\xi) \) denote the receiver’s informed optimal action given the actual value of \( \xi \). We have

\[ a_{NI} = \frac{r}{c} E[q] = \frac{r(\bar{\xi} + \xi)}{c}, \]
\[ a_I(\xi) = \frac{r}{c} E[q | \xi] = \frac{r}{c} \xi. \]

One can employ the concept of perfect Bayesian equilibrium to obtain a solution for this interaction. Özer et al. (2018) show that all equilibria in this cheap-talk game is neither informative nor influential. This result is a consequence of the sender being insatiable (i.e., sender’s payoff is strictly monotone in receiver’s action). Intuitively, regardless of her information \( \xi \), the sender would prefer to send the message(s) that will induce the highest action 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 uninformed action, ignoring the sender’s messages. Theorem 1 formally states this result. The proofs for all theorems are deferred to Appendix A.

**Theorem 1.** In any perfect Bayesian equilibrium, the sender’s equilibrium strategy \( \hat{\xi}(\xi) \) is uncorrelated with \( \xi \), and the receiver’s equilibrium strategy is \( a(\hat{\xi}) = a_{NI} \).

Contrary to these standard theory predictions, Özer et al. (2018) find in their experiments that
there is distorted, yet, informative and influential communication. Specifically, there is positive correlation between sender messages and sender private information, even though the messages are inflated relative to the private information. There is also positive correlation between sender messages and receiver actions, even though receiver actions are lower than what is optimal if the sender messages are taken to be truthful, i.e., receivers discount sender messages. Thus, there is over-communication relative to standard theory predictions. Moreover, there is considerable heterogeneity in the extent to which senders distort and receivers discount the message. We next turn to the behavioral economics models of cheap talk as two alternative models to explain these observed behaviors. We start with the level-k model.

3.2 Level-k Model and Predictions

Level-k models start with the specification of the $L_0$ player’s behavior, which is non-strategic and represents naive or unsophisticated play. The $L_0$ player’s behavior serves as the starting point for each higher type’s model of others since $L_1$ best responds to $L_0$, $L_2$ best responds to $L_1$ and so on. In cheap-talk games (such as the one we study), $L_0$ senders are specified as truthful, and $L_0$ receivers are specified as credulously believing the sender’s message to be truthful (Crawford 2003; Cai and Wang 2006; Kawagoe and Takizawa 2009; Ellingsen and Östling 2010; Wang et al. 2010). 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, even if he [subsequently] ends up not taking it at face value”\(^9\). Note that $L_0$ receivers (who are believing) are also effectively best responding to $L_0$ senders (who are truthful); hence, there is no need to separately consider that $L_1$ receivers best respond to $L_0$ senders ($L_0$ $L_1$ is $L_0$ $L_2$ best responds to $L_0$ $L_1$ and so on). Therefore, with some abuse of notation, we consider that $L_k$ receivers best respond to $L_k$ senders; $L_k$ senders still best respond to $L(k-1)$ receivers. Thus, $L_1$ senders best respond to $L_0$ receivers, $L_1$ receivers best respond to $L_1$ senders and so on.

For an $L_k$ sender (e.g., a supplier participant), let $\Pi_{Sk}(\hat{\xi}, \xi)$ denote the expected payoff given the actual state $\xi$ (e.g., her private demand information) and her message $\hat{\xi}$ (e.g., her demand report), and let $\hat{\xi}_k(\xi)$ denote her best response message given $\xi$. For an $L_k$ receiver (e.g., a retailer

\(^9\)In level-k models of other games, $L_0$ player’s behavior is often specified as a random action, i.e., $L_0$ player picks any action among possible actions with equal probability. However, this specification does not allow us to explain observed behaviors in cheap-talk games. Specifically, if the $L_0$ sender uniformly randomizes over all possible messages, then her messages are uninformative. Hence, the $L_1$ receiver will not be influenced by any message. In turn, the $L_2$ sender will be indifferent between sending any of the messages. Thus, the model up to this point cannot explain informative or influential communication and, without additional assumptions, cannot make further behavioral predictions for higher types. Assuming that the $L_2$ sender will be truthful (when indifferent between the messages), will lead to equivalent behavioral predictions for higher types as starting with a truthful $L_0$ player and obtaining the behaviors of the higher types.
participant), let $\Pi_{Rk}(a, \xi)$ denote the expected payoff given the sender’s message $\hat{\xi}$ and the receiver’s action $a$ (e.g., his service level), and $a_k(\hat{\xi})$ denote his best response to the sender’s message $\hat{\xi}$. The $L0$ sender’s behavior is $\hat{\xi}_0(\xi) = \xi$. The $L0$ receiver naively believes that $\xi = \hat{\xi}$, i.e., whatever $L0$ sender reports. Therefore, from Equation (4), $L0$ receiver’s best response to the message $\hat{\xi}$ is $a_0(\hat{\xi}) = a_1(\hat{\xi}) = \frac{\xi}{c} x^*$. We obtain the behaviors of higher level senders and receivers iteratively as follows.

An $L1$ sender believes the receiver is $L0$ who will take action $a_0(\hat{\xi})$ in response to a message $\hat{\xi}$. Hence, the $L1$ sender’s expected payoff is

$$
\Pi_{S1}(\hat{\xi}, \xi) = sa_0(\hat{\xi}) E[q \mid \xi] = \frac{s}{c} \xi \hat{\xi},
$$

which is strictly increasing in $\hat{\xi}$. Hence, the $L1$ sender’s best response is to send the message $\hat{\xi}_1(\xi) = \bar{\xi}$ for any $\xi$, so as to induce $L0$ receiver to choose the highest possible action (e.g., service level).

Next, consider the $L1$ receiver. The $L1$ receiver should expect that the $L1$ sender sends the message $\hat{\xi} = \bar{\xi}$ for any $\xi$, and hence this message is not informative. To determine the receiver’s behavior for any message $\hat{\xi} < \bar{\xi}$ (which cannot be from an $L1$ sender), we must specify any $Lk$ receiver’s belief following an “off-equilibrium” message. We assume that an $Lk$ receiver believes that an off-equilibrium message is from the next highest sender type (lower than $Lk$) that uses this message (as in Kawagoe and Takizawa 2009, and Ellingsen and Östling 2010).\footnote{Such a lower sender type always exists because the $L0$ sender’s strategy has full support, i.e., it sends all messages in the message space $[\xi, \bar{\xi}]$ with positive probability. So any message that is off-equilibrium for an $Lk$ sender is used by the $L0$ sender, and may also be used by a higher sender type between $L0$ and $Lk$.} This 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 imply that the receiver believes that the sender is of the appropriate lower type.\footnote{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. Specifically, the $L1$ receiver always takes the uninformed action $a_{NL}$ for all sender messages. Further, the $L2$ sender is then indifferent between sending any message, and the behavior of higher types cannot be fixed without additional assumptions. Also, no new sender or receiver behaviors are predicted compared to the approach we follow. 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 $L(k-1)$. Consequently, the set of behaviors predicted by this approach will in general change if $Lk$ player’s belief is even slightly perturbed to allow for arbitrarily small positive probability of lower type players.}

In particular, because behavior is derived iteratively, the

$\hat{\xi}$ $E$ $q$ $c$ $s$
Therefore, the Lk receiver essentially follows the strategy of the L(k − 1) receiver for any off-equilibrium message (Ellingsen and Östling 2010).

Therefore, in the case of the L1 receiver, a message \( \hat{\xi} \leq \bar{\xi} - 1 \) is taken to be from an L0 sender, and hence the receiver believes that \( \xi = \bar{\xi} \). A message \( \hat{\xi} = \bar{\xi} \) is taken to be from an L1 sender, hence uninformative, and the L1 receiver believes that \( \mathbf{E}[\xi \mid \hat{\xi} = \bar{\xi}] = \mathbf{E}[\xi] = \frac{\xi + \bar{\xi}}{2} \). Therefore, an L1 receiver’s expected payoff is
\[
\Pi_{R1}(a, \hat{\xi}) = \begin{cases} 
    r\hat{\xi}a - \frac{1}{2}ca^2, & \hat{\xi} \leq \bar{\xi} - 1; \\
    r\frac{\hat{\xi} + \bar{\xi}}{2}a - \frac{1}{2}ca^2, & \hat{\xi} = \bar{\xi}.
\end{cases}
\] (6)

Essentially, the L1 receiver believes all messages \( \hat{\xi} < \bar{\xi} \), and disregards the message \( \hat{\xi} = \bar{\xi} \). It follows that the L1 receiver’s best response is to take the action \( a_I(\hat{\xi}) \) if \( \hat{\xi} \leq \bar{\xi} - 1 \), and the uninformed action \( a_{NI} \) if \( \hat{\xi} = \bar{\xi} \). Therefore
\[
a_I(\hat{\xi}) = \begin{cases} 
    a_I(\bar{\xi}) = \frac{s}{c}\bar{\xi}, & \hat{\xi} \leq \bar{\xi} - 1; \\
    a_{NI} = \frac{s}{c}(\frac{\hat{\xi} + \bar{\xi}}{2}), & \hat{\xi} = \bar{\xi}.
\end{cases}
\] (7)

An L2 sender anticipates the above behavior of the L1 receiver. Hence, an L2 sender’s expected payoff is
\[
\Pi_{S2}(\hat{\xi}, \xi) = \begin{cases} 
    sa_I(\hat{\xi}) \xi = \frac{s}{c}\bar{\xi}\xi, & \hat{\xi} \leq \bar{\xi} - 1; \\
    sa_{NI}\xi = \frac{s}{c}(\frac{\hat{\xi} + \bar{\xi}}{2})\xi, & \hat{\xi} = \bar{\xi}.
\end{cases}
\] (8)

The L2 sender’s expected payoff is strictly increasing in \( \hat{\xi} \) for \( \hat{\xi} \leq \bar{\xi} - 1 \). In addition, for sufficiently large discrete message space such that \( \xi - 1 > \frac{\xi + \bar{\xi}}{2} \) (which is the case in the experiments we study), the sender’s expected payoff is higher for \( \hat{\xi} = \bar{\xi} - 1 \) than for \( \hat{\xi} = \bar{\xi} \) because \( a_I(\bar{\xi} - 1) > a_{NI} \). Therefore, for any \( \xi \), the L2 sender always sends the message \( \hat{\xi}_2(\xi) = \bar{\xi} - 1 \), so as to induce the highest action an L1 receiver would choose.

Next, the L2 receiver expects that the L2 sender sends the message \( \hat{\xi} = \bar{\xi} - 1 \), and hence this message is not informative. The message \( \hat{\xi} = \bar{\xi} \) is taken to be from the next highest type that uses the message, namely the L1 sender whose strategy is \( \hat{\xi}_1(\xi) = \bar{\xi} \). Hence, this message is also uninformative. Any message \( \hat{\xi} \leq \bar{\xi} - 2 \) is taken to be from the L0 sender and, thus, truthful. Therefore, the L2 receiver believes \( \xi = \hat{\xi} \) if \( \hat{\xi} \leq \bar{\xi} - 2 \), and \( \mathbf{E}[\xi \mid \hat{\xi}] = \frac{\xi + \bar{\xi}}{2} \) for \( \hat{\xi} > \bar{\xi} - 2 \). Accordingly, receiver must believe (with probability 1) that the sender is of type L0 since L0’s strategy has full support.
an L2 receiver’s expected payoff is

\[ \Pi_{R2}(a, \hat{\xi}) = \begin{cases} r\hat{\xi}a - \frac{1}{2}ca^2, & \hat{\xi} \leq \bar{\xi} - 2; \\ r\frac{(\hat{\xi} + \xi)}{2}a - \frac{1}{2}ca^2, & \hat{\xi} > \bar{\xi} - 2. \end{cases} \]  

(9)

It follows that the L2 receiver takes the action \( a_I(\hat{\xi}) \) if \( \hat{\xi} \leq \bar{\xi} - 2 \), and the uninformed action \( a_{NI} \) if \( \hat{\xi} > \bar{\xi} - 2 \). Therefore

\[ a_2(\hat{\xi}) = \begin{cases} a_I(\hat{\xi}) = \frac{\hat{\xi}}{c}\xi, & \hat{\xi} \leq \hat{\xi}_k; \\ a_{NI} = \frac{\xi(\hat{\xi} + \xi)}{2}, & \hat{\xi} > \hat{\xi}_k. \end{cases} \]  

(10)

Continuing iteratively in this manner, we obtain the behavior of all higher types. We find that, for any \( \xi \), an \( Lk \) sender sends the highest message that the \( L(k - 1) \) receiver will believe, namely \( \hat{\xi}_k(\xi) = \hat{\xi} - (k - 1) \), provided doing so would induce a higher action than \( a_{NI} \); if all messages that the receiver \( L(k - 1) \) 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 \( \hat{\xi}_k(\xi) = \frac{\xi + \xi}{2} \). Correspondingly, an \( Lk \) receiver takes the action \( a_I(\hat{\xi}) \) for all messages that are lower than the one sent by the \( Lk \) sender (i.e., \( \hat{\xi} \leq \hat{\xi}_k(\xi) - 1 \)), believing them to be from the truthful \( L0 \) sender; and takes the uninformed action \( a_{NI} \) for all other messages, believing them to be uninformative. The following theorem summarizes our results.

**Theorem 2.** For \( k > 0 \), an \( Lk \) sender’s strategy is \( \hat{\xi}_k(\xi) = \hat{\xi}_k = \max\{\bar{\xi} - (k - 1), \frac{\xi + \xi}{2}\} \), and an \( Lk \) receiver’s strategy is

\[ a_k(\hat{\xi}) = \begin{cases} a_I(\hat{\xi}) = \frac{\hat{\xi}}{c}\xi, & \hat{\xi} \leq \hat{\xi}_k; \\ a_{NI} = \frac{\xi(\hat{\xi} + \xi)}{2}, & \text{otherwise}, \end{cases} \]

where \( \hat{\xi}_k = \hat{\xi}_k - 1 \).

Thus, contrary to the standard theory model, the level-k model can potentially explain the over-communication phenomenon, as well as heterogeneity in sender and receiver behaviors. Communication is informative because \( L0 \) senders are truthful. Further, a higher-level \( (k > 0) \) sender distorts the message to a particular message level \( (\hat{\xi}_k) \), which depends on the sender’s level of thinking, so as to induce the highest action that an immediately lower-level receiver would choose following any message. Thus, for example, an \( L1 \) sender inflates to the maximum extent, an \( L2 \) sender inflates almost to the maximum, and so on. Hence, there is heterogeneity in the extent to which senders distort the message. Communication is influential since a receiver is influenced by sender messages up to a threshold message level \( (\hat{\xi}_k) \) that the receiver believes are not distorted. Further, the threshold message level depends on the receiver’s level of thinking. Thus, for example,
an $L1$ receiver believes all messages except the highest one, an $L2$ receiver believes all messages except the highest two messages and so on. Hence, there is also heterogeneity in the extent to which different receiver types are influenced by the message.

We further find that, in our setting with an insatiable sender, it is not necessary to separately consider the predictions of the cognitive hierarchy model. The cognitive hierarchy model differs from the level-$k$ model in that a higher-level player assigns positive probability to the other player being of any lower-level type as per a truncated Poisson distribution. We derive the predictions of the cognitive hierarchy model in Appendix B, and find that the predicted behavior of an $Lk$ player is the same (for sender) or practically the same (for receiver) as an $Lk$ or lower-level player in the level-$k$ model. Thus, essentially, no new behaviors are predicted. Further, the cognitive hierarchy model also predicts a narrower range of variation in behaviors across player types than the level-$k$ model. Therefore, we focus only on the level-$k$ model predictions in the rest of the paper.

### 3.3 Trust-embedded Model and Predictions

The trust-embedded model of cheap talk, first proposed and tested in Özer et al. (2011), is motivated by research on trust and trustworthiness. In the context of communication games, trust has been 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. Trustworthiness has been defined as the sender’s tendency not to manipulate her message to her own monetary benefit at the expense of the receiver (see also Özer and Zheng 2018 for a comprehensive discussion on the definition of trust and trustworthiness and related literature).

The trust-embedded model differs from the standard economic analysis of cheap talk in three respects. First, a sender can be trustworthy because she 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, and experience an intrinsic cost from lying (e.g., Gneezy 2005; Mazar et al. 2008; Lundquist et al. 2009; Erat and Gneezy 2012; Scheele et al. 2018). This lying cost can be attributed to factors such as adherence to social norms, maintaining positive self-image, avoiding hurting others, and other social motives. A common conclusion in this literature is that decision makers trade off the monetary payoff from lying with the intrinsic cost of lying in deciding whether and how much to lie. Furthermore, the cost of lying depends on the “size of the lie" (e.g., Gneezy et al. 2018; Scheele et al. 2018), with the marginal cost of a lie is increasing in the magnitude of a lie (e.g., Mazar et al. 2008; Kartik 2009).

Accordingly, the trust-embedded model assumes that a sender incurs a lying cost $G \left( \hat{\xi} - \xi ; \gamma \right)$
that is strictly convex and increasing in the magnitude of the lie $|\hat{\xi} - \xi|$, where $\gamma \geq 0$ denotes the sender’s lying cost type. Let $g \left( |\hat{\xi} - \xi|; \gamma \right) = \frac{\partial G \left( |\hat{\xi} - \xi|; \gamma \right)}{\partial \hat{\xi}}$ denote the marginal cost of lying. The lying cost type $\gamma$ indexes the lying cost such that: (i) a sender incurs no lying cost if $\gamma = 0$, i.e., $G \left( |\hat{\xi} - \xi|; \gamma = 0 \right) = 0$; and (ii) the sender’s marginal cost of lying $g \left( |\hat{\xi} - \xi|; \gamma \right)$ is strictly increasing in $\gamma$ for all $|\hat{\xi} - \xi| > 0$. We impose the regularity condition that $G \left( |\hat{\xi} - \xi|; \gamma \right)$ is twice continuously differentiable for $\xi, \hat{\xi} \in [\xi, \bar{\xi}]$; hence, in particular, $g \left( 0; \gamma \right) = 0$.

Second, the standard economic model assumes that a receiver follows Bayes rule to update his belief about $\xi$ given $\hat{\xi}$, anticipating the sender’s communication strategy for each $\xi$. 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). Similarly, Sheremeta and Shields (2013) and Jin et al. (2018) find that receivers assign higher subjective probabilities of senders’ messages being truthful than should be expected from a rational Bayesian decision maker. In fact, receivers are overly trusting even in games where standard equilibrium predicts that the receivers disregard any sender message (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). This tendency to trust is consistent with an “experiential view” of trust and trustworthiness that suggests that a person’s disposition for trust or trustworthiness in any situation is gradually formed through life experiences, influenced by social norms, environment, or values that a person adheres to, and unlikely to change in a single interaction (e.g., Moorman et al. 1993; Doney and Cannon 1997; Brehm and Rahn 1997; Hardin 2002; Özer et al. 2011; Beer et al. 2018).

Accordingly, the trust-embedded model assumes that a receiver follows a significantly simpler updating rule that reflects his inherent tendency to trust. This rule is represented in the form of the following non-Bayesian belief about $\xi$ conditional on $\hat{\xi}$: receivers believe that $\xi$ 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 $\xi^T$ follows the distribution of $\xi$ truncated on $[\xi, \hat{\xi}]$. Essentially, the receiver believes with probability $\alpha_R$ that $\hat{\xi}$ is completely truthful; and believes with probability $1 - \alpha_R$ that $\hat{\xi}$ is inflated but otherwise uninformative, i.e., $\hat{\xi}$ is inflated by some $\delta \in [\hat{\xi} - \xi, \hat{\xi}]$ that follows a uniform probability distribution.\footnote{Without loss of generality, we include $\delta = 0$ to keep the subsequent analytical expressions simple.}

$\alpha_R$ is interpreted as the receiver’s tendency to be trusting. If $\alpha_R = 1$, then the receiver fully trusts
that the sender is truthful and believes $\xi = \hat{\xi}$. If $\alpha_R = 0$, then the receiver does not trust the message at all, that is, he fully expects the sender to have inflated the message by some $\delta \in [0, \hat{\xi} - \xi]$ and believes that the true state is less than the received message, all values being equally likely. In other words, $\xi \sim Unif [\xi, \hat{\xi}]$, and the message is otherwise uninformative.

Lastly, the trust-embedded model assumes that the sender has a belief about the receiver’s trust type $\alpha_R$, denoted by $\alpha_S$ 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.

Applying the trust-embedded model to our setting, we obtain the following. Consider a receiver of trust type $\alpha_R$. On receiving a message $\hat{\xi}$, the receiver updates his belief according to the trust-based updating rule described above, and we have

$$
E \left[ \xi \mid \hat{\xi}, \alpha_R \right] = \alpha_R \hat{\xi} + (1 - \alpha_R)E \left[ \xi \mid \xi \leq \hat{\xi} \right],
$$

$$
= \alpha_R \hat{\xi} + (1 - \alpha_R)\frac{\hat{\xi} + \xi}{2} = \frac{(1 + \alpha_R)\hat{\xi} + (1 - \alpha_R)\xi}{2}.
$$

(11)

In particular, $E \left[ \xi \mid \hat{\xi}, \alpha_R \right] = \hat{\xi}$ if $\alpha_R = 1$ since the receiver fully trusts the message; and $E \left[ \xi \mid \hat{\xi}, \alpha_R \right] = \frac{\hat{\xi} + \xi}{2}$ if $\alpha_R = 0$ since the receiver believes $\xi \sim Unif [\hat{\xi}, \hat{\xi}]$. The receiver’s expected payoff is

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

$$
= ra \left[ \frac{(1 + \alpha_R)\hat{\xi} + (1 - \alpha_R)\xi}{2} \right] - \frac{1}{2}ca^2.
$$

(12)

Let $a \left( \hat{\xi}; \alpha_R \right)$ denote the receiver’s strategy. Then, the expected utility for a sender of lying cost type $\gamma$ and belief $\alpha_S$ about the receiver’s trust type is

$$
\Pi_S \left( \hat{\xi}, \xi; \gamma, \alpha_S \right) = sE \left[ a \left( \hat{\xi}; \alpha_R \right) \mid \alpha_S \right] \xi - G \left( \left| \hat{\xi} - \xi \right| ; \gamma \right),
$$

(13)

where we use $\Pi_S$ to denote utility (i.e., the monetary payoff minus the cost of lying). Let $\hat{\xi} \left( \xi; \gamma, \alpha_S \right)$ denote the sender’s strategy.

We derive the optimal sender and receiver strategies by backward induction from Equations (12) and (13). From Equation (11), 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 his action is increasing in the sender’s message. Accordingly, the sender can have an incentive to inflate the message depending on her lying cost. Theorem 3 formally describes the sender’s and receiver’s optimal strategies.
Theorem 3. The sender’s optimal strategy is
\[
\hat{\xi}^* (\xi; \gamma, \alpha_S) = \min \left\{ \xi, \xi + g^{-1} \left( s^T \left( \frac{1 + \alpha_S}{2} \right) \xi \right) \right\},
\]
where \( \bar{\alpha}_S = \int \alpha dH(\alpha) \). The receiver’s optimal strategy is
\[
a^* (\hat{\xi}; \alpha_R) = \frac{r}{c} \left[ \frac{(1 + \alpha_R) \hat{\xi} + (1 - \alpha_R) \xi}{2} \right].
\]

Theorem 3 leads to the following observations. First, the sender inflates the message by 
\( g^{-1} \left( s^T \left( \frac{1 + \alpha_S}{2} \right) \xi \right) \) or up to \( \hat{\xi} \), whichever is lower. The extent to which the sender inflates the message is (weakly) decreasing in the lying cost type \( \gamma \), since the marginal cost of lying \( g(\cdot) \) is increasing in \( \gamma \). In the empirical application, we use a quadratic lying cost function 
\( G\left( \hat{\xi} - \xi \mid \gamma \right) = \frac{\gamma}{2} (\hat{\xi} - \xi)^2 \), and the marginal lying cost is 
\( g\left( \hat{\xi} - \xi \mid \gamma \right) = \gamma |\hat{\xi} - \xi| \). Therefore, for \( \gamma \to \infty \), lying becomes unattractive and the sender’s strategy converges to truth-telling. i.e., she is fully trustworthy. Whereas, for \( \gamma = 0 \), the sender always inflates up to the maximum extent, i.e., she is fully untrustworthy. For \( \gamma \) between these two extremes, the sender has an incentive to inflate the message by an intermediate amount.

Note that the sender’s trustworthiness also depends on his belief \( \bar{\alpha}_S \) about the receiver’s trust type. All else equal, the sender has a higher incentive to inflate the message if she expects the receiver’s trust type to be higher. Therefore, both \( \gamma \) and \( \bar{\alpha}_S \) jointly determine the extent to which the sender inflates the message and, thus, her trustworthiness. Specifically, for the aforementioned quadratic lying cost, we have from Theorem 3, that 
\( \hat{\xi}^* (\xi; \gamma, \alpha_s) = \min \left\{ \xi, \xi \cdot A(\gamma, \bar{\alpha}_S) \right\} \), where 
\( A(\gamma, \bar{\alpha}_S) = 1 + \frac{1}{\gamma} s^T \left( \frac{1 + \alpha_S}{2} \right) \). We refer to \( A(\gamma, \bar{\alpha}_S) \) as the sender’s trustworthiness factor, since it captures how much the sender inflates her message; a higher trustworthiness factor indicates lower trustworthiness since the sender inflates the message more.

Next, turning to the receiver’s strategy, the extent to which the receiver is influenced by the message is increasing in his trust type \( \alpha_R \). If \( \alpha_R = 1 \), then the receiver trusts the sender’s message and his optimal strategy is 
\( a\left( \hat{\xi}; \alpha_R = 1 \right) = a_f \left( \hat{\xi} \right) = \frac{\hat{\xi}}{\xi} \). If \( \alpha_R = 0 \), then the receiver’s optimal strategy is 
\( a\left( \hat{\xi}; \alpha_R = 0 \right) = \frac{\hat{\xi}}{\xi} + \frac{\xi}{\xi - \hat{\xi}} \). Note that the receiver is still influenced even if \( \alpha_R = 0 \) because he believes that the true state 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 can also potentially explain participants’ observed behaviors in the experiments, such as the over-communication phenomenon, as well as heterogeneity in sender and receiver behaviors. Sender messages are inflated, yet informative if the lying cost is sufficiently high, and receivers discount sender messages and are yet influenced according to the trust-based
inference rule. Moreover, there is heterogeneity in the extent to which the sender distorts and the receiver is influenced depending on their trustworthiness factor and trust type, respectively.

3.4 Comparing the Predictions of the Behavioral Models

Both the level-k model as well as the trust-embedded model predict distorted yet informative and influential communication. Further, both models can potentially explain the over-communication phenomenon and the heterogeneity in sender and receiver behaviors. However, the two models predict characteristically distinct individual-level behaviors for senders and receivers. In particular, these differences arise because the sender is insatiable.

In the case of senders, the level-k model assumes senders are motivated by their pecuniary interests but limited in their strategic ability, i.e., level of thinking. As a result, the level-k model predicts a higher-level sender inflates the message up to a level (i.e., $\hat{\xi}_k$) that depends on the sender’s level of thinking, irrespective of her private information. In contrast, the trust-embedded model assumes senders are motivated by their pecuniary interest, but are also curbed by their trustworthiness, i.e., lying cost. As a result, the trust-embedded model predicts a trustworthy sender inflates her message to a level that is curbed by her lying cost, which depends on her private information. This dependence allows us to identify how sender’s trustworthiness constrains her incentive to distort the message (for different values of her private information).

In the case of receivers, the level-k model assumes receivers follow Bayesian inference based on iterative levels of strategic thinking about senders. As a result, the level-k model predicts a level-k receiver believes all messages up to a threshold ($\tilde{\xi}_k$), and are hence influenced by them. Yet, he also believes any higher than this threshold to be distorted, and is not influenced by it. In contrast, the trust-embedded model assumes receivers follow a trust-based (non-Bayesian) inferences rule depending on their trust factor. As a result, the trust-embedded model predicts that a trusting receiver need not completely believe or disbelieve a sender’s message, and may be (partially) influenced by all messages even if the receiver believes that the sender may have distorted her message.

Thus, for the cheap-talk game in Özer et al. (2018), the two behavioral models have two distinct ways through which they explain over-communication phenomenon and heterogeneity in behaviors. The large strategy space in this cheap-talk game further allows for these differences to be empirically identified. Hence, we can let the experimental data inform us as to which of the two models better explain observed behaviors.
4 Data and Empirical Methodology for Model Comparison

In this section, we summarize the experimental data, the structural estimation approach, and the metrics we use to compare how well each model predicts participants’ observed behavior.

4.1 Experimental Data

Özer et al. (2018) conduct five experimental sessions - three main sessions and two additional sessions with the following experimental parameters: \( \xi \sim \text{Unif}[10, 80] \), \( \epsilon \sim \text{Unif}[-10, 10] \), \( a \in [0, 180] \), \( s = \frac{1}{2} \), \( r = 1 \), \( c = 1 \). Therefore, \( \Pi_S = \frac{1}{2}aq \), \( \Pi_R = aq - \frac{1}{2}a^2 \) in Equations (1) and (2); and \( a_{NI} = 45 \), \( a_I(\xi) = \xi \leq 80 \) in Equations (3) and (4). Each experimental session consisted of 12 participants and 11 decision rounds. The value of \( \xi \) and \( \epsilon \) varied randomly from one decision round to the next. Participants were paired randomly and anonymously in each decision round and never rematched to minimize repeated interaction effects. To ensure familiarity with both roles, participants were rotated between the sender and the receiver roles from one decision round to the next. Thus, each participant completed five decisions in one role and six decisions in the other.\(^{14}\) The two additional sessions were conducted to examine whether and how encouraging more analytical thinking by senders affected sender and receiver decisions. 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. The authors find that standard theory predictions are rejected by the experiment: in the main sessions, sender’s messages are significantly correlated with their private information \( (\rho = 0.66, p < 0.01) \), and the receiver’s actions are significantly correlated to the sender’s message \( (\rho = 0.36, p < 0.01) \).\(^{15}\) Furthermore, compared to the main sessions, sender participants inflated their messages significantly more and receiver participants discount sender messages significantly more in the additional sessions.\(^{16}\) As explained later in §4.4, we leverage the additional sessions to determine whether the behavioral models are able to recover the effect of the experimental manipulation.

---

\(^{14}\)Participants played two unpaid practice rounds (one round in each role) prior to the 11 paid decision rounds.

\(^{15}\)All p-values are two-sided unless mentioned otherwise.

\(^{16}\)Özer et al. (2018) define the extent to which the sender inflates the message as \( \hat{\xi} - \xi \), and the extent to which the sender message is discounted as the difference \( a - a_I(\hat{\xi}) \) between the receiver’s observed action and the receiver’s optimal action if the sender’s message were taken to be truthful. The level of inflation in the main sessions (mean 11.8, median 5) was significantly lower than in the additional sessions (mean 16.6, median 10, \( p < 0.01 \) under Kruskal Wallis test), and the level of discounting in the main session was also lower in the main sessions (mean 17.8, median 12.5) than in the additional sessions (mean 19.9, median 17.5, \( p = 0.08 \) under Kruskal Wallis test).
4.2 Model Estimation Approach

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 Iribarri 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 theoretically predicted payoff (under a particular model) and a logit shock. The player chooses the decision that yields the highest utility. The logit shock can cause the player to deviate from his or her theoretically predicted decision.\footnote{Wang et al. (2010) find that in their experiments (with small strategy spaces), participants often chose a theoretically predicted action. Therefore, the 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 Iribarri (2007), in our context (with large strategy spaces), participants often do not choose a theoretically predicted action and we adopt the standard logit formulation.} Nevertheless, the theoretically predicted decision has a higher likelihood of being chosen and, hence, being observed in the data.\footnote{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. We follow the above approach to be consistent with the literature on level-k models.}

To capture the heterogeneous behaviors of different types of players, we allow for a mixture of player types and estimate the type probabilities from the data. All model parameters are estimated using maximum likelihood. As the model parameters for senders and for receivers are separable, we estimate the models separately for senders and for receivers to evaluate how well each model performs in explaining the behavior in each role.

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 a receiver ($R$). Let $\Omega_P$ denote the set of player types for role $P$, and $\pi_\omega$ denote the probability of player type $\omega \in \Omega_P$ such that $\sum_{\omega \in \Omega_P} \pi_\omega = 1$. In the level-k model, a player type refers to the level of thinking. In the trust-embedded model, player type refers to the trust or trustworthiness types. 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 $\xi_t$ 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 $\{I_{it}, D_{it}\} = \{\xi_t, \hat{\xi}_{it}\}$ for a participant in the sender role and $\{I_{it}, D_{it}\} = \{\hat{\xi}_{it}, a_{it}\}$ for a participant in the receiver role. Let $\Delta_P$ denote the set of feasible decisions in role $P$; $\Delta_S$ is the set of integers in $[\xi, \bar{\xi}]$ and $\Delta_R$ is the set of integers in $[0, \bar{a}]$. Let $\Pi_P(D_{it}, I_{it}; \theta_\omega)$ denote participant $i$’s expected 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 $\theta_\omega$ denotes type specific...
parameters in the payoff function. In the level-k model, \( \Pi_P \) is given by Equations (20) and (21). In the trust-embedded model, \( \Pi_P \) is given by Equations (12) and (13).

Under the logit random-utility formulation, conditional on participant \( i \) in role \( P \) being of type \( \omega \in \Omega_P \), the probability that the participant makes decision \( D_{it} \in \Delta_P \) given information \( I_{it} \) in round \( t \in T_{iP} \) is

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

where \( \lambda_\omega \) 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 will uniformly randomize across all actions.

Therefore, the likelihood of observing decisions \( D_{iP} = \{D_{it} : t \in T_{iP}\} \) by participant \( i \) in role \( P \) given information \( I_{iP} = \{I_{it} : t \in T_{iP}\} \) unconditional on type is

\[
L_{iP}(D_{iP} \mid I_{iP}, \pi^P, \lambda^P, \theta^P) = \sum_{\omega \in \Omega_P} \prod_{t \in T_{iP}} Pr_P(D_{it} \mid I_{it}, \theta_\omega, \lambda_\omega),
\]

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 log-likelihood function for model parameters for role \( P \) given observations \( D^P = \{D_{iP} : i \in N\} \) and \( I^P = \{I_{iP} : i \in N\} \) across all participants is

\[
LL_P(\pi^P, \lambda^P, \theta^P | D^P, I^P) = \sum_{i \in N} \log \left( L_{iP}(D_{iP} \mid I_{iP}, \pi^P, \lambda^P) \right).
\]

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

### 4.3 Player Types in Level-k and Trust-embedded Models

**Level-k Model.** Prior empirical applications of the level-k model across a wide variety of games have found that player types higher than \( L_3 \) are rare, with most players being of type \( L_1 \) or \( L_2 \) (e.g., see surveys by Crawford et al. 2013; Georganas et al. 2015). As shown in §3.2, in our context, the payoffs and predicted behaviors of level \( k > 0 \) senders and receivers change very gradually. Consequently, the behaviors of \( L_1 \), \( L_2 \) and \( L_3 \) types are too similar to be empirically distinguishable. Also, the \( L_0 \) receiver’s behavior (who essentially best responds to the \( L_0 \) sender) is also very close to the that of \( L_1 \), \( L_2 \) and \( L_3 \) receivers. Therefore, we estimate one \( L_1 \sim 3 \) sender type based on the behavior of the \( L_1 \) sender, and one \( L_0 \sim 3 \) receiver type based on the behavior of the \( L_1 \) receiver. As in Wang et al. (2010), we include an \( L_0 \) sender type who deviates from her specified truthful behavior as per a noise distribution; specifically, we assume a truncated normal shock, whose variance \( \sigma_{L0}^2 \) is estimated from the data. We follow this approach because, unlike in the

24
case of the other player types, the (truthful) behavior of the $L_0$ sender cannot be obtained through payoff maximization (for any belief about the receiver’s strategy).

In addition to the above types, we allow for a higher type (of sender and receiver), denoted as $LH$, whose level of thinking $H$ is estimated from the data. This approach also allows us to handle the possibility that behaviors change in larger steps than the gradual change predicted by the level-k model. Specifically, from Theorem 2 in §3.2, an $LH$ sender will send a message $\hat{\xi}_{LH} = \tilde{\xi} - (H - 1)$ if $H \leq \frac{\tilde{\xi} + \xi}{2}$ (= 45), and $\hat{\xi}_{LH} = \frac{\tilde{\xi} + \xi}{2}$ otherwise. An $LH$ receiver fully believes all messages less than a threshold $\tilde{\xi}_{LH}$ and ignores any higher message, where $\tilde{\xi}_{LH} = \tilde{\xi} - H$ if $H \leq \frac{\tilde{\xi} + \xi}{2}$ and $\tilde{\xi}_{LH} = \frac{\tilde{\xi} + \xi}{2}$ otherwise. Further, the predicted payoffs for a $LH$ sender and receiver can be obtained from Equations (20) and (21), respectively. We estimate the $LH$ sender’s message $\hat{\xi}_{LH}$ and the $LH$ receiver’s threshold $\tilde{\xi}_{LH}$ from the data, allowing the estimated levels of thinking to be different for senders and receivers.

Thus, we adopt a more flexible level-k structure than commonly used in prior applications, thereby allowing the level-k model more flexibility to explain observed behavior in the data. We further remark that some level-k model applications for non-cheap-talk games omit the $L_0$ player from model estimation, essentially assuming that this player type exists only in the minds of players as a starting point for boundedly-strategic thinking. In particular, in their application of the level-k model to auctions, Crawford and Iribarri (2007) start with a truthful $L_0$ type to obtain level-k model predictions, but do not include this type in their model estimation. For our purpose of comparing the two behavioral models, we include the $L_0$ type to give the level-k model sufficient room to explain observed behaviors. Later, in §6, we discuss the implications of omitting the truthful $L_0$ type for model estimation.

**Trust-Embedded Model.** We consider a high ($H$) and low ($L$) trustworthiness type for senders, and a high ($H$) and low ($L$) trusting type for receivers. Allowing for additional types does not substantially improve model performance. We remark that in the trust-embedded model, the logit error precision cannot be uniquely determined for senders, because their utility can be rescaled with respect to the logit error precision $\lambda_\omega$ without affecting the logit choice probabilities. Specifically, given $\xi$, the sender’s logit utility (with shock $\epsilon$) from sending message $\hat{\xi}$ is

$$U_S\left(\hat{\xi}, \xi; \gamma_\omega, \bar{\alpha}_{S_\omega}, \lambda_\omega\right) = s \left(\frac{r}{c} \left[\frac{(1-\bar{\alpha}_{S_\omega})\hat{\xi} + (1-\bar{\alpha}_{S_\omega})\hat{\xi}_S}{2}\right]\right) \xi - \frac{1}{2} \gamma_\omega \left(\hat{\xi} - \xi\right)^2 + \frac{1}{\lambda_\omega} \epsilon.$$

$$= C + A(\gamma_\omega, \bar{\alpha}_{S_\omega}) \xi - \frac{1}{2} \gamma_\omega \hat{\xi} + \frac{1}{\lambda_\omega} \epsilon,$$

where $C = s \frac{r}{c} \frac{(1-\bar{\alpha}_{S_\omega})\xi}{2} - \frac{1}{2} \gamma_\omega \xi^2$ does not depend on the message choice $\hat{\xi}$, and hence does not
Consequently, the trustworthiness factor $A(\gamma_\omega, \bar{\alpha}_s\omega)$ and the scaled lying cost $\lambda_\omega \gamma_\omega$ (which jointly affect the logit probabilities) are uniquely identified. For receivers, the trust type $\alpha_\omega$ and the logit error precision $\lambda_\omega$ are both identified.

### 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 an experimental manipulation using the data from 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:

1. **Mean Squared Error** (MSE) between the observed and predicted behaviors; a lower mean square error indicates better forecasting performance,
2. **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 reliably determine out-of-sample forecasting performance with limited data, we use Monte Carlo Cross-Sample Validation (MCCV) (Hastie et al. 2011). The essential idea of MCCV is as follows. 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 sampling process 1000 times and calculate the average forecasting performance across these samples. 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.

To determine a model’s 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. Recall that in the additional sessions, participants in the sender role were required to make payoff calculations that highlighted the benefit

\footnote{Using Equation (17), for senders in the trust embedded model, the logit choice probability in Equation (14) can be written as

$$
Pr_S(\xi_{it} \mid \xi_i, \gamma_\omega, \alpha_s\omega, \lambda_\omega) = \frac{\exp\{\lambda_\omega \gamma_\omega F(\hat{\xi}_{it}, \hat{\xi}; A(\gamma_\omega, \bar{\alpha}_s\omega))\}}{\sum_{\xi \in \Delta_S} \exp\{\lambda_\omega \cdot \gamma_\omega \cdot F(\hat{\xi}, \hat{\xi}; A(\gamma_\omega, \bar{\alpha}_s\omega))\}},
$$

where $F(\hat{\xi}, \xi; A(\gamma_\omega, \bar{\alpha}_s\omega)) = A(\gamma_\omega, \bar{\alpha}_s\omega) \cdot \xi \hat{\xi} - \frac{1}{2} \xi^2$.}
of inflating their message before they made their decisions. The level-k theory would predict that this manipulation would cause more participants to engage in higher levels of thinking. The trust-embedded model theory would predict that participants exhibit lower trust and trustworthiness because the manipulation makes the pecuniary payoff more salient, and can thus crowd out non-pecuniary motives. We determine whether the estimated model parameters in the additional sessions are consistent with these predictions for each model.

5 Which Behavioral Economics Model Performs Better?

5.1 Model Comparison for Senders

Table 1 summarizes the estimation results for the level-k and trust-embedded models for senders. The level-k model estimates suggest that a substantial portion of participants are of type $L_0$ (47%, 17 out of 36) sending truthful messages, and a substantial proportion are of type $L_1 \sim 3$ (47%, 17 out of 36) inflating the message close to the maximum extent possible. The remaining are of type $L_H$ (6%, 2 out of 36) distorting the message to a message level $\hat{\xi}_{LH} = 51$, which suggests a relatively high level of thinking ($H = 30$). The trust-embedded model estimates suggest that the high trustworthy type is close to being fully-trustworthy or truthful, inflating the message by $A_H = 1.03$ times (which is not significantly different than 1). The low trustworthy type, in contrast, inflates the message considerably, specifically by $A_L = 2.35$ times (which is significantly different than 1), yet not always to the maximum extent (which would require a trustworthiness factor of 8). A majority of participants are classified to be of low trustworthy type (56%, 20 out of 36).

Comparing the in-sample model fits of both models, the trust-embedded model performs better, with lower AIC and BIC. We discuss possible reasons later in §5.3. Further, the trust-embedded model considerably outperforms the level-k model in out-of-sample forecasting accuracy, with much lower prediction errors ($MSE$) and higher goodness of fit ($\hat{\beta}, R^2$). 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 §4.4).

Recall from §4.1 that sender participants inflated messages considerably more in the additional sessions than in the main sessions. Under the level-k model, we find that the level of thinking of the $LH$ type is practically the same in the main and additional sessions. However, there is an increase in the proportion of $L_1 \sim 3$ and in the proportion of $LH$ types at the expense of the $L_0$ type.

---

$^3$Participants are classified to be of the type for which their estimated posterior probability of being a particular type is highest.

$^4$For all significance results of model estimates, we construct 95% confidence intervals using bootstrapping.
(47.22% vs. 25%, \( p = 0.04 \)) 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. Similarly, 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 before), both of which translate to higher inflation in sender messages. 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 |
|-----------------|-----------------|
| **Model Estimation** | **Classification** |
| | \( L_0 \) (47.22%) | \( L_1 \sim 3 \) (47.22%) | \( LH \) (5.56%) | \( High \) (16 (44.44%)) | \( Low \) (20 (55.56%)) |
| **Model Parameters** | \( \xi_k \) | 51.00* | \( A_H \) | 1.03* |
| | \( \sigma_{L_0} \) | 5.84 | \( A_L \) | 2.35* |
| | \( \lambda_{L_1 \sim 3} \) | 1.59* | \( \lambda_H \cdot \gamma_H \) | 33.35* |
| | \( \lambda_H \) | 3.26* | \( \lambda_L \cdot \gamma_L \) | 0.47 |
| **In-Sample Model Fit** | \( LL \) | -715.77 | -709.58 |
| | \( AIC \) | 1443.54 | 1429.16 |
| | \( BIC \) | 1463.27 | 1445.60 |
| **MCCV** | \( MSE \) | 461.70 | 281.72 |
| | \( \hat{\beta} \) | 0.61 | 0.83 |
| | \( R^2 \) | 0.33 | 0.52 |
| **Experimental Manipulation** | **Classification** |
| | \( L_0 \) (25.00%) | \( L_1 \sim 3 \) (58.33%) | \( LH \) (16.67%) | \( High \) (7 (29.17%)) | \( Low \) (17 (70.83%)) |
| **Model Parameters** | \( \xi_k \) | 52.00* | \( A_H \) | 1.17* |
| | \( \sigma_{L_0} \) | 8.22* | \( A_L \) | 6.31* |
| | \( \lambda_{L_1 \sim 3} \) | 2.51* | \( \lambda_H \cdot \gamma_H \) | 23.10* |
| | \( \lambda_H \) | 2.11* | \( \lambda_L \cdot \gamma_L \) | 0.15 |
| **LL** | -478.84 | -471.75 |
| **AIC** | 969.68 | 953.50 |
| **BIC** | 986.98 | 967.91 |

* significant at 95% level; we use bootstrapping to determine the significance level.

5.2 Model Comparison for Receivers

Table 2 summarizes the estimation results for receivers. The level-k model estimates suggest that all participants are of type \( L_0 \sim 3 \), and are thus believing of most messages except possibly extreme ones (close to 80). For an \( LH \) receiver, the estimated threshold up to which they believe messages \( \bar{\xi}_{LH} = 65 \), which suggests a relatively high level of thinking \( H = 16 \). However, none of the participants are classified to be of this type, i.e., the posterior probability of being this type is less
than 0.5 for all participants. The trust-embedded model estimates suggest that while a significant proportion of participants (42%, 15 out of 36) are of high trust type and are highly influenced by the message ($\alpha_{RH} = 0.72$, not significantly less than 1), the remaining low trust participants discount the message considerably ($\alpha_{RL} = 0$, significantly less than 1).²²

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

<table>
<thead>
<tr>
<th>Model Estimation</th>
<th>Classification</th>
<th>Level-K Model</th>
<th>Trust-embedded Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L0 \sim 3$</td>
<td>$LH$</td>
<td>$High$</td>
</tr>
<tr>
<td></td>
<td>$L0$</td>
<td>$36$ (100%)</td>
<td>$15$ (41.67%)</td>
</tr>
<tr>
<td></td>
<td>$LH$</td>
<td>0 (0%)</td>
<td>$21$ (58.33%)</td>
</tr>
<tr>
<td>Model Parameters</td>
<td>$\xi_{LH}$</td>
<td>65.01*</td>
<td>$\alpha_{RH}$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{L0 \sim 3}$</td>
<td>0.61*</td>
<td>$\alpha_{RL}$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{LH}$</td>
<td>6.00*</td>
<td>$\lambda_{H}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\lambda_{L}$</td>
</tr>
<tr>
<td>In-Sample Model Fit</td>
<td>LL</td>
<td>-840.31</td>
<td>-758.82</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>1688.62</td>
<td>1527.64</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>1701.77</td>
<td>1544.08</td>
</tr>
<tr>
<td>MCCV</td>
<td>MSE</td>
<td>288.27</td>
<td>113.36</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.46</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.17</td>
<td>0.62</td>
</tr>
<tr>
<td>Experimental Manipulation</td>
<td>Classification</td>
<td>$L0 \sim 3$</td>
<td>$High$</td>
</tr>
<tr>
<td></td>
<td>$LH$</td>
<td>22 (91.67%)</td>
<td>3 (12.50%)</td>
</tr>
<tr>
<td></td>
<td>$L0$</td>
<td>2 (8.33%)</td>
<td>21 (87.50%)</td>
</tr>
<tr>
<td>Model Parameters</td>
<td>$\xi_{LH}$</td>
<td>50.08*</td>
<td>$\alpha_{RH}$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{L0 \sim 3}$</td>
<td>9.50e-11</td>
<td>$\alpha_{RL}$</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{LH}$</td>
<td>12.00</td>
<td>$\lambda_{H}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\lambda_{L}$</td>
</tr>
<tr>
<td>In-Sample Model Fit</td>
<td>LL</td>
<td>-560.81</td>
<td>-520.25</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>1129.62</td>
<td>1050.50</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>1142.77</td>
<td>1066.94</td>
</tr>
</tbody>
</table>

* significant at 95% level; we use bootstrapping to determine the significance level.

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 considerably outperforms the level-k model in out-of-sample forecasting accuracy, with much lower prediction errors ($MSE$) and higher goodness of fit ($\hat{\beta}$, $R^2$). Finally, the trust-embedded model recovers the effect of the experimental manipulation better. Recall from §4.1 that receiver participants discounted sender messages considerably more in the additional sessions than in the main sessions. The level-k model classifies almost all participants as type $L0 \sim 3$ in the additional sessions, and thus indicates that the experimental manipulation did not have much effect on participants’ strategic level of thinking (and, hence, the extent to which they discounted the message). In contrast, the trust-embedded model finds a higher proportion of low-trust type in the additional sessions than in

²²Low trust receivers are still influenced by the sender’s message $\xi$ because they believe that the true state is less than the received message. Specifically, from Theorem 3 in §3.3, $a^*(\xi; \alpha_R = 0) = \xi \mathbb{E} \left[ \xi \mid \xi < \hat{\xi} \right] = \frac{1}{2} (\hat{\xi} + \xi)$. 

29
the main sessions (87.5% vs. 58.33%, p < 0.05), consistent with receivers discounting the message to a greater extent. The estimated trust parameter of the low-trust type ($\alpha_{RL}$) is higher than in the main sessions, though this does not per-se contradict lowering of trust since more participants are of low trust. The trust-embedded model also has better in-sample fit than the level-k model in the additional sessions, with considerably lower $AIC$ and $BIC$.

5.3 Discussion

Overall, our results indicate that the trust-embedded model performs better than the level-k model in explaining and predicting human participants’ actions in a cheap-talk context. In comparing the explanatory powers of the estimated models for senders, we observe that the truthful $L0$ sender behavior in the level-k model is very similar to the estimated behavior of the high trustworthy type in the trust-embedded model. However, the behaviors of the $L1 \sim 3$ and $LH$ senders in the level-k model differ considerably from that of the low trustworthy type in the trust-embedded model. Moreover, the messages of participants classified as $L1 \sim 3$ and $LH$ senders in the level-k model are significantly correlated with their private information ($\rho = 0.53$, $p < 0.01$), which is contrary to the prediction of the level-k model (which predicts that they distort the message to a particular message level and, hence, uncorrelated with their private information). In contrast, the trust-embedded model predicts significant correlation between messages and actual information for the low trustworthy type senders ($\rho = 0.79$ between $\xi$ and predicted $\hat{\xi} = \max\{80, A_L\xi\}$), and the observed correlation is significant ($\rho = 0.53$, $p < 0.01$). Thus, the trust-embedded model better captures the behaviors of senders who send inflated, yet informative messages.

In comparing the explanatory powers of the estimated models for receivers, we observe that the level-k model classifies all participants to be of type $L0 \sim 3$, who are fully-influenced by all sender messages except some extreme ones. Now, the estimated behavior of the high trust type in the trust-embedded model is not significantly different than fully-trusting behavior, and hence somewhat similar to the $L0 \sim 3$ type in the level-k model. However, in the trust-embedded model, a majority of participants are classified to be of the low trust type, who significantly discount sender messages and are hence only partially influenced by sender messages. Thus, the trust-embedded model performs better by capturing the behaviors of receivers who are only partially influenced by messages.

The above insights are robust to augmenting the level-k model to include player types who follow standard theory predictions. Even though the level-k model posits that human players are boundedly sophisticated in their strategic thinking, prior research has found that some participants
may yet behave consistent with standard equilibrium predictions. In Appendix E, we show that including such player types does not improve model performance for senders; the model performance for receivers is improved but is still considerably poorer than the trust-embedded model.

6 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. The level-k model takes the grim view that human decision makers are only self-interested, but are limited in their ability to think strategically. The trust-embedded model takes a more positive view that decision makers are not only guided by their self-interest, but also by non-pecuniary motives to be trusting and trustworthy. We compare the two models using a cheap-talk context and experiments in which the two models yield characteristically distinct predictions and let the data inform us which model performs better. Our work represents a step towards understanding what aspects of human behavior are relatively more important to explain how cheap talk works in practice. These results also pave the way for richer and more realistic models of cheap talk that can better predict outcomes in business and economic settings.

Comparing the Two Contrasting Perspectives: Overall, the results provide evidence for the more positive perspective of human behavior in cheap-talk interactions, namely, that decision makers are guided by non-pecuniary motives to be trusting and trustworthy. Thus, effective cheap-talk communication is not simply a matter of boundedly-strategic thinking in the sole pursuit of self-interest. The present study shows that the trust-embedded model explains sender as well as receiver behaviors better than the level-k model in cheap-talk experiments with an insatiable sender. Specifically, by explicitly considering that senders have an intrinsic lying cost, the trust-embedded model is able to account for not only truthful (fully-trustworthy) communication but also partially distorted but informative communication (that is not predicted by the level-k model). Moreover, by allowing for a non-Bayesian inference process that reflects a receiver’s innate tendency to trust, the trust-embedded model is able to account for receivers who are moderately influenced by sender messages (whose behavior cannot be explained by the level-k model that assumes a Bayesian inference process based on iterative levels of strategic thinking). From a managerial perspective, these results suggest that to improve transactions based on cheap talk, firms should focus on designing processes to reduce barriers for and to engineer trusting and trustworthy relationships, for example, by reducing perceived vulnerabilities and social uncertainties in market environment, understanding cultural difference across countries and institutions, establishing mutually trusting and
trustworthy relationship through intra- and inter-organizational networks (Özer and Zheng 2018), or through suitable reputation and feedback systems (Bolton et al. 2013).

**Common Context to Compare the Models:** Our work also provides insights about developing a common context to compare models of boundedly-strategic behavior and social preferences. As noted by Sobel (2013), comparing models of boundedly-strategic thinking and non-pecuniary motives is an important gap in the current understanding of cheap talk. At the same time, distinguishing between these behavioral theories can be challenging because both theories can lead to qualitatively similar predictions. For example, 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 level-k model predicts inflated yet informative communication by higher-level senders, and that receivers are 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 §3 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. 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 that a cheap-talk context in which the sender is insatiable and the strategy spaces are large is well-suited to conduct further comparisons of the two models.

In this regard, we examined whether other previous cheap-talk experiments also meet these two criteria. We categorize prior cheap-talk experiments based on these two criteria in Appendix C.23 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. To the best of our knowledge, only Özer et al. (2018) use experiments both with an insatiable sender and large

---

23We 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 (Spiliotopoulos et al. 2016) and the receiver’s strategy may not depend on the sender’s message (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).
strategy spaces.\textsuperscript{24}

**Boundaries to the Applicability of Level-k Models:** Our findings also add to the discussion regarding the applicability of level-k models. In examining the generalizability of level-k thinking across games, Georganas et al. (2015) find that the level of strategic sophistication of individual participants is not persistent across different families of games, and even within the same family of games in some cases. They conjecture that level-k play may be triggered mainly in normal-form (simultaneous move) games with complete information. In contrast, a cheap-talk game we study is an extensive-form game with incomplete information. Furthermore, Heap et al. (2014) find that assumptions about $L_0$ behavior made outside the model for specific applications are not “portable” (or generalizable) across comparable normal-form two-player games. They conclude that the ad-hoc approach of tailoring $L_0$ type’s non-strategic behavior to suit a particular application of the level-k model can result in a high incidence of false positives for the explanatory power of the level-k model. Our findings add to this concern. In cheap-talk games, researchers have fixed the behavior of the $L_0$ player (outside of the model) to be truthful for senders, since doing so can explain observed behaviors. We find that a cheap-talk game with an insatiable sender puts the assumption about $L_0$ behavior more directly to the test. In this case, the level-k model predicts that higher level senders always distort the message to a particular message level. The level-k model also predicts specific behaviors for receivers (fully believing messages up to a threshold level and fully disbelieving all other messages) that are anchored in the assumed behavior of the $L_0$ sender. As discussed above, these level-k predictions fare poorly in accounting for observed behaviors in this context.

Moreover, the presence and assumed behavior of the $L_0$ sender is in and of itself crucial for the explanatory power of the level-k model in our context, since only the presence of the $L_0$ sender can explain the occurrence of informative communication. In fact, the proportion of participants classified as type $L_0$ senders is 39% in the main session and 25% in the additional sessions, which are both surprisingly high compared to past level-k applications. Wang et al. (2010) find that only

\textsuperscript{24}We remark that Özer et al. (2011) and Özer et al. (2014) also use experiments with relatively large strategy spaces for senders and receivers. However, the sender’s payoff in their context is not strictly monotone for the entire range of receiver’s possible actions. As a result, under the level-k model, a higher-level sender’s message can be distorted yet informative, as the sender does not need to induce the highest action that a lower-level receiver is willing to take. We conjecture that distinguishing the two behavioral models in this case would be more difficult relative to the 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 that allows us to examine how well the two behavioral models capture the effect of the manipulation.
1 out of 15 (7%) sender participants in their cheap talk experiment was classified as $L0$. Cai and Wang (2006), using a rudimentary procedure to classify participants in their cheap-talk experiment, find that of the 24 participants who could be classified into a level-k type for senders, only 2 (8%) were classified as an $L0$ sender. More generally, prior empirical applications of the level-k model in other games have frequently found few or no $L0$ type players (Crawford et al., 2013; Georganas et al., 2015).\footnote{We remark that in applications of the cognitive hierarchy model, the frequency of $L0$ types is not separately estimated. Instead, a Poisson distribution is assumed for player types and only the mean of this distribution is estimated. However, level-k models estimated with the same data (e.g., beauty contest experiments) typically reveal few $L0$ type players.} However, this is not the case in our application. Moreover, we show in Appendix D that omitting the $L0$ sender type in estimating the level-k model substantially worsens the level-k model’s performance. These observations add to the concern regarding the applicability of the level-k model in this context.

**In Closing:** We note that these results represent a concrete step towards understanding what behavioral drivers are important to capture and explain observed behaviors in a cheap-talk context that well represents many practical situations. Our findings encourage us to call for, in particular, two types of follow-on research: to better understand in other business and economic environments which of these two aspects of human behavior (bounded rationality and propensity to trust and be trustworthy) mostly guide managerial decisions, and how these two aspects can jointly influence business outcomes. More specifically, one direction for further research is to compare the level-k and trust-embedded models in different business interactions and contexts. For example, Özer et al. (2018) also consider an advice-giving game in which the sender can not only distort the advice in his or her favor to benefit at the receiver’s expense, but also, unlike in the information-sharing game, provide advice that can never be optimal for the receiver for any underlying state of the world. 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. In addition, future research may also consider developing cheap-talk experiments satisfying the two criteria (insatiable sender and large strategy spaces) to conduct further comparisons of the two behavioral theories. A second direction is to pursue enhanced analytical models that can capture both aspects of human behavior simultaneously. For example, in response to concerns about the generalizability of the level-k model, Crawford (2014) suggests that, rather than rejecting level-k model, the way forward could be to identify a robust hybrid of level-k thinking with “other kinds of thinking”. As noted in §5.3, augmenting the level-k model to include player types that behave according to standard theory predictions does not sufficiently improve level-k model performance. We conjecture that bringing
trust and trustworthiness (and related analytical models) into level-k thinking would likely make the refined level-k model more predictive.

References


Appendix A  Proofs for Theorems

Proof for Theorem 1: Let \( \hat{\xi}(\xi) \) denote the sender’s equilibrium strategy. Then, the receiver’s optimal strategy is

\[
a(\hat{\xi}) = \arg \max_{\xi} raE\left[\xi | \hat{\xi}\right] - \frac{1}{2}ca^2 = \frac{rE[\xi | \hat{\xi}]}{c}.
\]

Let \( a_{max} = \max_{\xi} a(\hat{\xi}) \) be the highest action that the receiver will take for any sender message. Let \( \Xi_{max} = \arg \max_{\xi} a(\hat{\xi}) \) be the set of messages that induce \( a_{max} \). Since the sender’s preference is strictly monotone in the receiver’s action, the sender will only send messages in \( \Xi \subseteq \Xi_{max} \) so as to induce the highest action. Moreover, the sender’s strategy must be such that the receiver finds it optimal to take the same action \( a_{max} \) for all messages in \( \Xi \). Therefore, \( E[\xi | \hat{\xi}] \) is the same for all \( \hat{\xi} \in \Xi \) in equilibrium. For the retailer’s inference to be unbiased in equilibrium, we require that

\[
E_{\hat{\xi}}\left[ E[\xi | \hat{\xi}] \right] = E[\xi] \quad \text{for} \quad \hat{\xi} \in \Xi.
\]

Therefore, \( E[\xi | \hat{\xi}] = E[\xi] \) for all \( \hat{\xi} \in \Xi \); otherwise, the retailer’s inference is biased. Consequently, \( a(\hat{\xi}) = aNI \) for all \( \hat{\xi} \) used by the sender in equilibrium. To see that \( \hat{\xi} \) is uncorrelated with \( \xi \) in any equilibrium, note that

\[
\text{Cov}(\hat{\xi}, \xi) = E[\hat{\xi}\xi] - E[\hat{\xi}]E[\xi] = E_{\hat{\xi}}\left[ E[\hat{\xi}\xi | \hat{\xi}] \right] - E[\hat{\xi}]E[\xi] = 0,
\]

where we have substituted \( E[\xi | \hat{\xi}] = E[\xi] \).

Proof for Theorem 2: We prove this result inductively. Suppose an \( L(k-1) \) receiver believes messages \( \hat{\xi} \leq \hat{\xi} - k + 1 \) and takes action \( a_I(\hat{\xi}) \), and disregards all higher messages \( \hat{\xi} > \hat{\xi} - k + 1 \) and takes action \( a_{NI} \). We showed that this holds for \( L(k-1) = L2 \) receiver in the main text. Now, an \( Lk \) sender’s expected payoff is

\[
\Pi_{Sk}(\hat{\xi}, \xi) = \begin{cases} 
sa_I(\hat{\xi}) \xi = sI^2 \hat{\xi} \xi, & \hat{\xi} \leq \hat{\xi} - k + 1; \\
sa_{NI} \xi = sI^2 \left( \hat{\xi} + \xi \right) / 2 \xi, & \hat{\xi} > \hat{\xi} - k + 1.
\end{cases}
\]

Consequently, regardless of his or her true information \( \xi \), the \( Lk \) sender sends the highest message that the \( L(k-1) \) receiver will believe, namely \( \hat{\xi}_k(\xi) = \hat{\xi} - (k-1) \), 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 \( \hat{\xi}_k(\xi) = \hat{\xi} + \xi / 2 \). Therefore, \( \hat{\xi}_k(\xi) = \hat{\xi}_k = \max\left\{ \hat{\xi} - (k-1), \frac{\hat{\xi} + \xi}{2} \right\} \). An \( Lk \) receiver will believe all messages \( \xi \leq \hat{\xi}_k = \hat{\xi}_k - 1 \), taking them to be from an \( L0 \) sender; and will disregard all higher messages \( \xi > \hat{\xi}_k \), taking them to be from the appropriate sender type from \( L1 \) to \( Lk \), and hence
uninformative. Therefore the \( L_k \) receiver’s expected payoff is

\[
\Pi_{R_k}(a, \hat{\xi}) = \begin{cases} 
  r\xi a - \frac{1}{2}ca^2, & \hat{\xi} \leq \hat{\xi}_k; \\
  \frac{r(\hat{\xi} + \xi)}{2} - a - \frac{1}{2}ca^2, & \hat{\xi} > \hat{\xi}_k
\end{cases}
\] (21)

Hence, the \( L_k \) receiver takes the action \( a_I(\hat{\xi}) \) for \( \hat{\xi} \in [\xi, \hat{\xi}_k] \), and the action \( a_{NI} \) otherwise.

**Proof for Theorem 3:** Maximizing the receiver’s expected payoff in Equation (12), we obtain \( a^*(\hat{\xi}; \alpha_R) \) as stated in the theorem. For the sender, given the belief \( \alpha_S \) about the receiver’s trust type, we have

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

Therefore, from Equation (13), the sender’s expected utility is

\[
\Pi_S(\hat{\xi}, \xi; \gamma, \alpha_s) = \frac{s}{c} \left[ \frac{(1 + \bar{\alpha_S})\hat{\xi} + (1 - \bar{\alpha_S})\xi}{2} \right] \xi - G\left(\hat{\xi} - \xi \mid \gamma\right).
\] (23)

which is strictly increasing in \( \hat{\xi} \) for \( \hat{\xi} \leq \xi \), and strictly concave in \( \hat{\xi} \) for \( \hat{\xi} > \xi \). Hence, \( \Pi_S(\hat{\xi}, \xi; \gamma, \alpha_s) \) is unimodal and there is a unique optimal message \( \hat{\xi}^*(\xi; \gamma, \alpha_S) \geq \xi \). The first order condition for \( \hat{\xi}^*(\xi; \gamma, \alpha_S) \) from Equation (23) yields

\[
\frac{\partial \Pi_S(\hat{\xi}^*(\xi; \gamma, \alpha_S); \xi; \gamma, \alpha_s)}{\partial \hat{\xi}} = \frac{s}{c} \left( \frac{1 + \bar{\alpha_S}}{2} \right) \xi - g(\hat{\xi}^*(\xi; \gamma, \alpha_S) - \xi; \gamma).
\] (24)

Since \( g(0; \gamma) = 0 \), it follows that \( \hat{\xi}^*(\xi; \gamma, \alpha_S) = \min \left\{ \hat{\xi}, \xi + g^{-1}\left( \frac{s}{c} \frac{1 + \bar{\alpha_S}}{2} \xi \right) \right\} \).

**Appendix B  Cognitive Hierarchy (CH) Model Predictions**

In this section, we develop the CH model and its predictions. We compare these predictions with those from the level-k model to show that (a) the CH model does not predict new behaviors for senders than the level-k model; (b) the CH model does not predict qualitatively new behaviors for receivers than the level-k model, and the difference in predicted behaviors is not empirically distinguishable; (c) the CH model predicts a narrower range of possible behaviors across types than the level-k model.

The CH model assumes that an \( L_k \) player believes that other player(s) could be of any type from \( L_0 \) through \( L(k - 1) \), with the probability of a particular type following a truncated Poisson distribution over these types and the mean \( \tau \) of the (non-truncated) Poisson distribution being the same for all \( k > 0 \). Since the predictions of the CH model will depend on \( \tau \), where necessary, we describe the results for specific values of \( \tau \) in the range \([0.25, 5]\). Past empirical applications of the CH model across a wide variety of games show that \( \tau \) is typically in the range \([1, 2]\).
We start by specifying the $L_0$ type behaviors. As in the level-k model, the $L_0$ sender is truthful, reporting $\hat{\xi}_0(\xi) = \xi$, and the $L_0$ receiver is credulous, taking action $a_0(\hat{\xi}) = a_I(\hat{\xi})$. An $L_k$ sender believes that the receiver could be of any type from $L_0$ through $L_{(k-1)}$, following a truncated Poisson distribution over types $t \in \{0, 1 \ldots k - 1\}$ with non-truncated mean $\tau$. An $L_k$ receiver believes that the sender could be of any type from $L_0$ through $L_k$, following a truncated Poisson distribution over types $t \in \{0, 1 \ldots k\}$ with non-truncated mean $\tau$. Let $p_t(\tau) = \frac{e^{-\tau} \tau^t}{t!}$ denote the Poisson distribution probability for outcome $t \in \{0, 1, 2 \ldots\}$.

The $L_k$ receiver also updates his belief about the sender’s type in a Bayesian manner based on the sender’s message, correctly anticipating the strategies of sender types $L_0$ through $L_k$. Let $f_t(\hat{\xi})$ denote the probability that a type $t$ sender sends the message $\hat{\xi}$, and $e_t(\hat{\xi})$ denote the expected value of $\xi$ conditional on the message $\hat{\xi}$ from a type $t$ sender. Then, the $L_k$ receiver’s expected value of $\xi$ conditional on the message $\hat{\xi}$ is

$$E^k[\xi | \hat{\xi}] = \frac{\sum_{t=0}^{k} e_t(\hat{\xi}) f_t(\hat{\xi}) p_t(\tau)}{\sum_{t=0}^{k} f_t(\hat{\xi}) p_t(\tau)}.$$  \hfill (25)

Therefore, the $L_k$ receiver’s expected payoff from action $a$ is

$$\Pi_{Rk}(a; \hat{\xi}) = r E^k[\xi | \hat{\xi}] a - \frac{1}{2} c a^2.$$  \hfill (26)

Therefore, his optimal action

$$a_k(\hat{\xi}) = \frac{r}{c} E^k[\xi | \hat{\xi}].$$  \hfill (27)

Next, consider the $L_k$ sender. She correctly anticipates the response $a_t(\hat{\xi})$ of receiver types $t \in \{0, 1 \ldots k - 1\}$. Let $E^k[a_t(\hat{\xi})]$ be the expected receiver action, given by

$$E^k[a_t(\hat{\xi})] = \frac{\sum_{t=0}^{k} a_k(\hat{\xi}) p_t(\tau)}{\sum_{t=0}^{k} p_t(\tau)}.$$  \hfill (28)

Then, the $L_k$ sender’s expected payoff from message $\hat{\xi}$ is

$$\Pi_{Sk} = s E^k[a_t(\hat{\xi})] \xi.$$  \hfill (29)

Therefore, the sender’s optimal message is

$$\hat{\xi}_k(\xi) \in \arg \max_{\hat{\xi}} E^k[a_t(\hat{\xi})].$$  \hfill (30)

We now derive the predicted behaviors for each type iteratively and compare these predictions with those from the level-k model. We start with a $L_1$ sender. The $L_1$ sender believes the receiver is type $L_0$ and hence sends the message $\hat{\xi}_1(\xi) = \bar{\xi} = 80$, which is the same as in the level-k model.
Next, consider a $L1$ receiver. He believes that the sender could be of type $L1$ or $L0$, assigning a probability $\frac{p_0}{p_0 + p_1}$ to type $L0$. Further, the $L1$ receiver updates her belief based on the sender’s message $\hat{\xi}$. Because the $L0$ sender type truthfully sends message in the range $[10, 80]$, all messages are equally probable, and the probability of receiving a message $\hat{\xi}$ from a $L0$ sender is $f_0(\hat{\xi}) = \frac{1}{71}$. Furthermore, the expected value of $\xi$ given $\hat{\xi}$ (from an $L0$ sender) is $e_0(\hat{\xi}) = \hat{\xi}$. Because the $L1$ sender type always sends the message $80$, $f_1(\hat{\xi}) = 0$ for $\hat{\xi} \leq 79$ and $f_1(80) = 1$. Further, the expected value of $\xi$ given $\hat{\xi} = 80$ is $e_1(80) = \frac{\hat{\xi} + 71}{2} = 45$.

Therefore, as in the level-$k$ model, if the $L1$ receiver receives a message $\hat{\xi} \leq 79$, then he updates his belief that the sender is of type $L0$ and, therefore, believes the message is truthful. In particular, we have $E^k[\xi \mid \hat{\xi}] = \hat{\xi}$. If he receives a message $\hat{\xi} = 80$, then the sender could be either $L0$ and $L1$, though the types differ in the probability with which they could have sent this message. From Equation (25), we have

$$E^1[\xi \mid \hat{\xi} = 80] = \frac{e_0(80)f_0(80)p_0(\tau) + e_1(80)f_1(80)p_1(\tau)}{f_0(80)p_0(\tau) + f_1(80)p_1(\tau)} = \frac{80 + 3195\tau}{1 + 71\tau}. \quad (31)$$

From Equation (27), the $L1$ receiver’s optimal action is

$$a_1(\hat{\xi}) = \begin{cases} \hat{\xi}, & \hat{\xi} \leq 79, \\ E^1[\xi \mid \hat{\xi} = 80], & \hat{\xi} = 80, \end{cases} \quad (32)$$

where we have substituted $\frac{\tau}{c} = 1$ for the cheap talk experiment. Thus, the $L1$ receiver’s behavior is the same as in the level-$k$ model for $\hat{\xi} \leq 79$, believing the message to be truthful. For $\hat{\xi} = 80$, the level-$k$ model predicts that the receiver ignores the message and takes action $a_1(80) = 45$ because the sender is assumed to be type $L1$. In contrast, the CH model in general predicts a higher action because the sender could still be of type $L0$ with finite positive probability. While the predicted action is $80$ if $\tau = 0$, it is decreasing in $\tau$, and converges quite rapidly to $45$: the predicted action is $46.9$ for $\tau = 0.25$, $45.5$ for $\tau = 1$ and $45.2$ for $\tau = 2$, and $45.1$ for $\tau = 5$. Thus, the predictions are practically the same for both models for reasonable values of $\tau$, i.e., $\tau \in [1, 2]$. Intuitively, even though the prior probability of the sender being type $L0$ may not be negligible, the posterior probability that the message $\hat{\xi} = 80$ is from the $L0$ type is considerably small for reasonable $\tau$.

Next, consider a $L2$ sender. She believes that the receiver can be of type $L0$ or $L1$, assigning a probability $\frac{p_0}{p_0 + p_1}$ to type $L0$. For messages $\hat{\xi} \leq 79$, both receiver types believe the message and take action $\hat{\xi}$. Therefore, the sender must at least inflate the message to 79. For $\hat{\xi} = 80$, the
$L_0$ receiver takes action 80, whereas the $L_1$ receiver takes the action $E^1[\xi | \hat{\xi} = 80]$. Hence, the expected receiver action is

$$E^1[a_t(80)] = \frac{a_0(80)p_0 + a_1(80)p_1}{p_0 + p_1} = \frac{80 + E^1[\xi | \hat{\xi} = 80]}{1 + \tau}$$

$$= \frac{80 + 5750\tau + 3195\tau^2}{1 + 72\tau + 71\tau^2}. \tag{33}$$

While the expected receiver action is 80 if $\tau = 0$, it is decreasing in $\tau$ and converging quite rapidly to 45: it is 73.4 for $\tau = 0.25$, 62.7 for $\tau = 1$, 56.8 for $\tau = 2$ and 50.9 for $\tau = 5$. Therefore, for reasonable values of $\tau$, the $L_2$ sender sends the message $\hat{\xi} = 79$, the same as the $L_2$ sender in the level-k model. We proceed similarly to obtain the predicted behaviors for higher types.

Table 3: Predictions of Level-k vs. CH model for Senders

<table>
<thead>
<tr>
<th>Model</th>
<th>$\tau$</th>
<th>Sender Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$L_1$</td>
</tr>
<tr>
<td>Level-k</td>
<td>-</td>
<td>80</td>
</tr>
<tr>
<td>CH</td>
<td>0.25</td>
<td>80</td>
</tr>
<tr>
<td>CH</td>
<td>0.5</td>
<td>80</td>
</tr>
<tr>
<td>CH</td>
<td>1</td>
<td>80</td>
</tr>
<tr>
<td>CH</td>
<td>1.5</td>
<td>80</td>
</tr>
<tr>
<td>CH</td>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>CH</td>
<td>5</td>
<td>80</td>
</tr>
</tbody>
</table>

Table 3 provides a comparison of predicted behaviors of sender types under the level-k and CH models. Specifically, it shows the message that will be sent by the $L_1$ to $L_6$ sender types. For the CH model, the predictions are shown for specific values of $\tau$ in the range $[0.25, 5]$. We observe that an $L_k$ sender in the CH model distorts the message to a particular message level that is the same as that sent by an $L_k$ or a lower level sender in the level-k model. The reason is that the CH model assigns higher probability to the lower receiver types than the level-k model. Moreover, the predicted behaviors in the CH model change over a narrower range with the player’s type than in the level-k model.

For example, Table 3 shows that, depending on $\tau$, an $L_5$ sender always sends the message 76, 77, 78 or 79, thus resembling the behavior, respectively, of an $L_2$, $L_3$, $L_4$ or $L_5$ sender in the level-k model. Further, for $\tau = 1$, all sender types higher than $L_3$ send same message 77; for $\tau = 2$, it can be shown that all sender types higher than $L_5$ send the same message 75. Intuitively, because the CH model assigns higher probability to lower types, especially for high $k$ and low $\tau$, the behaviors of the higher level senders in the CH model resembles that of lower level senders in the level-k model beyond a threshold level of thinking.
Table 4: Predictions of Level-k vs. CH model for Receivers

<table>
<thead>
<tr>
<th>Receiver Type</th>
<th>Level-k Model</th>
<th>CH Model ($\tau = 0.25$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sender Message ($\hat{\xi}$)</td>
<td>Sender Message ($\hat{\xi}$)</td>
</tr>
<tr>
<td></td>
<td>≤ 74</td>
<td>75</td>
</tr>
<tr>
<td>L1</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L2</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L3</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L4</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L5</td>
<td>$\hat{\xi}$</td>
<td>45</td>
</tr>
<tr>
<td>L6</td>
<td>$\hat{\xi}$</td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Receiver Type</th>
<th>CH Model ($\tau = 1$)</th>
<th>CH Model ($\tau = 1.5$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sender Message ($\hat{\xi}$)</td>
<td>Sender Message ($\hat{\xi}$)</td>
</tr>
<tr>
<td></td>
<td>≤ 74</td>
<td>75</td>
</tr>
<tr>
<td>L1</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L2</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L3</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L4</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L5</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L6</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Receiver Type</th>
<th>CH Model ($\tau = 2$)</th>
<th>CH Model ($\tau = 5$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sender Message ($\hat{\xi}$)</td>
<td>Sender Message ($\hat{\xi}$)</td>
</tr>
<tr>
<td></td>
<td>≤ 74</td>
<td>75</td>
</tr>
<tr>
<td>L1</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L2</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L3</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L4</td>
<td>$\hat{\xi}$</td>
<td>75</td>
</tr>
<tr>
<td>L5</td>
<td>$\hat{\xi}$</td>
<td>49.1</td>
</tr>
<tr>
<td>L6</td>
<td>$\hat{\xi}$</td>
<td>49.1</td>
</tr>
</tbody>
</table>

Table 4 provides a comparison of predicted behaviors of receiver types $L_1$ to $L_6$ under the level-k and CH models. Specifically, it shows the response of receivers to sender messages $\hat{\xi} \in [10, 80]$. For the CH model, the predictions are shown for specific values of $\tau$ in the range $[0.25, 5]$. We observe that the predictions of the CH model resemble those of level-k model in the following respects. First, a receiver believes all messages up to a threshold message level and then discounts all higher messages. Second, the threshold message level for an $L_k$ receiver in the CH model is the same as that of an $L_k$ or lower level receiver in the level-k model. Similar to the case of senders, the
reason is that the CH model assigns higher probability to the lower sender types than the level-k model. Lastly, the predicted actions for messages higher than the threshold in the CH model is practically the same as that of the corresponding receiver type with the same threshold in the level-k model; in particular, the predicted behaviors differ only for a few messages and are hence practically indistinguishable. Moreover, the predicted behaviors in the CH model change over a narrower range with the player’s type than in the level-k model.

For example, an $L_2$ receiver in the CH model believes all messages up to 78 (same as the $L_2$ receiver in the level-k model) and takes practically the same action for higher messages. Further, depending on $\tau$, an $L_5$ receiver believes all messages up to 78, 77, 76 or 75, and the behavior is practically indistinguishable from that of the $L_2$, $L_3$, $L_4$ or $L_5$ receiver, respectively, in the level-k model. Lastly, we note that for $\tau = 1$, all receiver types higher than $L_3$ believe messages up to 76 and discount messages 77 and higher; for $\tau = 2$, it can be shown that all receiver types higher than $L_5$ believe messages up to 74 and discount messages 75 and higher. Thus, the range of predicted behaviors is narrower than in the level-k model.

Appendix C Experimental Setup in Prior Cheap Talk Experiments

<table>
<thead>
<tr>
<th>Insatiable Sender</th>
<th># of Sender Messages</th>
<th># of Receiver Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dickhaut et al. (1995)</td>
<td>No</td>
<td>4</td>
</tr>
<tr>
<td>Forsythe et al. (1999)</td>
<td>Yes</td>
<td>3</td>
</tr>
<tr>
<td>Blume et al. (2001)</td>
<td>No</td>
<td>3</td>
</tr>
<tr>
<td>Sánchez-Pagés and Vorsatz (2007)</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Kawagoe and Takizawa (2009)</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Lundquist et al. (2009)</td>
<td>No</td>
<td>100</td>
</tr>
<tr>
<td>Sánchez-Pagés and Vorsatz (2009)</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Wang et al. (2010)</td>
<td>No</td>
<td>5</td>
</tr>
<tr>
<td>Özer et al. (2011)</td>
<td>No</td>
<td>301</td>
</tr>
<tr>
<td>Sheremeta and Shields (2013)</td>
<td>Yes</td>
<td>2</td>
</tr>
<tr>
<td>Özer et al. (2014)</td>
<td>No</td>
<td>301</td>
</tr>
<tr>
<td>Özer et al. (2018)</td>
<td>Yes</td>
<td>71</td>
</tr>
</tbody>
</table>
Appendix D  Estimating Level-k Model without L0 Sender

We estimate the level-k model for senders omitting the ad-hoc L0 sender type. Table 5 summarizes the estimation results. The estimates suggest that a majority of participants are of type \(L1 \sim 3\) (83%, 30 out of 36), and thus inflate the message close to the maximum extent possible. For an \(LH\) sender, the estimated uninformative message (\(\hat{\xi}_{LH} \approx 57\)) suggests a relative high level of thinking (\(H = 24\)). Without the L0 sender type, the level-k model performs much worse than trust-embedded model in out-of-sample forecasting accuracy, with much higher \(MSE\) and much lower \(\hat{\beta}\) and \(R^2\). The level-k model appears to directionally recover the effect of the experimental manipulation. There is an increase in the proportion of participants classified as \(LH\) in the additional sessions compared to the main sessions (33.33% vs. 16.67%, \(p = 0.13\)), and a slight increase in the estimated level of thinking for the \(LH\) type in the additional sessions (\(\hat{\xi}_{LH} = 52\), hence \(H = 29\)). Again, we observe much worse in-sample fit than both level-k model in §5 and the trust-embedded model, with considerably higher \(AIC\) and \(BIC\).

Table 5: Estimation Results of Level-k Model for Senders without L0 Sender

<table>
<thead>
<tr>
<th>Model Estimation</th>
<th>Classification</th>
<th>(L1 \sim 3)</th>
<th>(LH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Parameters</td>
<td>(\xi_{LH})</td>
<td>56.99*</td>
<td>6 (16.67%)</td>
</tr>
<tr>
<td></td>
<td>(\lambda_{L1 \sim 3})</td>
<td>0.94*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\lambda_{LH})</td>
<td>1.78*</td>
<td></td>
</tr>
<tr>
<td>In-Sample Model Fit</td>
<td>LL</td>
<td>-791.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>1591.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>1604.99</td>
<td></td>
</tr>
<tr>
<td>MCCV</td>
<td>MSE</td>
<td>978.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\hat{\beta})</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(R^2)</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>Experimental Manipulation</td>
<td>Classification</td>
<td>(L1 \sim 3)</td>
<td>(LH)</td>
</tr>
<tr>
<td>Model Parameters</td>
<td>(\xi_{LH})</td>
<td>52.00*</td>
<td>8 (33.33%)</td>
</tr>
<tr>
<td></td>
<td>(\lambda_{L1 \sim 3})</td>
<td>1.60*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\lambda_{LH})</td>
<td>1.78*</td>
<td></td>
</tr>
<tr>
<td>Model Fit</td>
<td>LL</td>
<td>-499.52</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>1007.04</td>
<td></td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>1020.19</td>
<td></td>
</tr>
</tbody>
</table>

* significant at 95% level; we use bootstrapping to determine the significance level.

Appendix E  Augmenting Level-k Model with Equilibrium Types

We verify the robustness of our findings regarding the level-k model by augmenting it with an additional \(EQ\) type for senders and receivers. For senders, Theorem 1 predicts that the message
is uncorrelated with their true information. Therefore, we include an EQ sender who uniformly randomizes the message. Note that while an uninformative sender may also chose to send the same message regardless of the true information, such a behavior resembles that of the LH sender whom we already include in the level-k model. For receivers, Theorem 1 predicts that their actions are not influenced by the message, and are based on the prior belief about the distribution of $\xi$. Therefore, we include an EQ receiver who maintains his prior belief about $\xi$.

Table 6: Augmented Level-k Model Estimation Results

<table>
<thead>
<tr>
<th>Model Estimation</th>
<th>Senders</th>
<th></th>
<th>Receivers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td></td>
<td>L0</td>
<td>16 (44.44%)</td>
<td>L0 $\sim$ 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L1 $\sim$ 3</td>
<td>15 (41.67%)</td>
<td>LH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LH</td>
<td>2 (5.56%)</td>
<td>EQ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EQ</td>
<td>3 (8.33%)</td>
<td></td>
</tr>
<tr>
<td>Model Parameters</td>
<td>$\xi_k$</td>
<td>55.99*</td>
<td>$\xi_k$</td>
<td>52.72*</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{L0}$</td>
<td>5.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\lambda_{L1-3}$</td>
<td>1.48*</td>
<td>$\lambda_{L0-3}$</td>
<td>6.61*</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{LH}$</td>
<td>4.65*</td>
<td>$\lambda_{LH}$</td>
<td>4.98*</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{EQ}$</td>
<td>-</td>
<td>$\lambda_{EQ}$</td>
<td>1.07*</td>
</tr>
<tr>
<td>In-Sample Model Fit</td>
<td>LL</td>
<td>-713.87</td>
<td>LL</td>
<td>-822.87</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>1441.74</td>
<td>AIC</td>
<td>1657.74</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>1464.76</td>
<td>BIC</td>
<td>1677.47</td>
</tr>
<tr>
<td>MCCV</td>
<td>MSE</td>
<td>392.04</td>
<td>MSE</td>
<td>118.71</td>
</tr>
<tr>
<td></td>
<td>$\bar{\beta}$</td>
<td>0.60</td>
<td>$\bar{\beta}$</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.28</td>
<td>$R^2$</td>
<td>0.32</td>
</tr>
<tr>
<td>Experimental Manipulation</td>
<td>Classification</td>
<td>L0</td>
<td>6 (25.00%)</td>
<td>L0 $\sim$ 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L1 $\sim$ 3</td>
<td>14 (54.17%)</td>
<td>LH</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LH</td>
<td>3 (16.67%)</td>
<td>EQ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EQ</td>
<td>1 (4.17%)</td>
<td></td>
</tr>
<tr>
<td>Model Parameters</td>
<td>$\xi_k$</td>
<td>50.24*</td>
<td>$\xi_k$</td>
<td>51.33*</td>
</tr>
<tr>
<td></td>
<td>$\sigma_{L0}$</td>
<td>8.21*</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$\lambda_{L1-3}$</td>
<td>2.652*</td>
<td>$\lambda_{L0-3}$</td>
<td>12.42*</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{LH}$</td>
<td>2.04*</td>
<td>$\lambda_{LH}$</td>
<td>12.80*</td>
</tr>
<tr>
<td></td>
<td>$\lambda_{EQ}$</td>
<td>-</td>
<td>$\lambda_{EQ}$</td>
<td>2.13e-07</td>
</tr>
<tr>
<td>Model Fit</td>
<td>LL</td>
<td>-479.20</td>
<td>LL</td>
<td>-560.80</td>
</tr>
<tr>
<td></td>
<td>AIC</td>
<td>972.40</td>
<td>AIC</td>
<td>1133.60</td>
</tr>
<tr>
<td></td>
<td>BIC</td>
<td>992.58</td>
<td>BIC</td>
<td>1150.90</td>
</tr>
</tbody>
</table>

* significant at 95% level; we use bootstrapping to determine the significance level.

Table 6 summarizes the estimation results. For senders, including the EQ sender does not improve the level-k model performance significantly with respect to in-sample fit, out-of-sample predictive performance or recovering the effect of an experimental manipulation. For receivers, including the EQ receiver substantially improves the level-k model performance with respect to in-
sample fit and out-of-sample predictive performance. Instead of all receivers being classified as $L0 \sim 3$, 75% of participants (27 of 36) are now classified as type $EQ$. We remark that prior applications of level-k models that have included an $EQ$ type have found that majority of participants are of type $L1$ or $L2$ (e.g., Costa-Gomes and Crawford (2006); Wang et al. (2010)). We suggest that in our case the high proportion of $EQ$ type is because the level-k model predictions for $L1$ or $L2$ fare poorly in accounting for observed behaviors. We also observe that there is an increase in the proportion of $EQ$ receiver in the manipulation sessions compared to the main sessions (91.67% vs. 75%, $p = 0.10$), which is consistent with participants thinking more deeply. Nevertheless, this augmented receiver model with $EQ$ type still does not perform as well as the trust-embedded model on these measures. Essentially, the augmented level-k model still does not allow for intermediate levels of influence of the message. Consequently, the trust-embedded model, which allows for varying level of trusting behaviors, better explains the observed interactions.