Team Decision Making in Operations Management

Jiawei Li, Damian R. Beil, Stephen G. Leider
Stephen M. Ross School of Business, University of Michigan, Ann Arbor, Michigan 48109
{jiawli@umich.edu, dbeil@umich.edu,leider@umich.edu}

The existing behavioral OM literature has primarily studied individual decision makers. However, the behavioral economics literature suggests that, compared to individuals, teams may make better decisions in tactical settings and may be more strategic and self-interested. To study the behavior of teams making operational decisions, we study two canonical settings: standalone Newsvendor inventory decisions (tactical decision-making) and Newsvendor under information sharing (strategic decision-making). We find that teams perform worse than individuals when making Newsvendor decisions, and exhibit a stronger pull-to-center bias. In the information sharing setting, team retailers are less trustworthy while team suppliers are just as trusting as individual suppliers. We provide evidence for a team-decision making framework that can organize these findings. To do so we leverage a novel aspect of our experiment: subject teams can text chat to facilitate decision making, and by analyzing team chats, we are able to explore the drivers of operational decisions in our setting. We also test various existing behavioral theories of Newsvendor decision making using our data, finding support for a modified version of mean anchoring and adjustment.

Key words: Team, Decision Making, Newsvendor, Information Sharing, Chat Analysis

1. Introduction

Business decisions are often made in a team setting. In a large consumer goods firm we interact with, the managers reported to us that at one large factory they have a dedicated scheduling team managing production, a different team responsible for determining the forecast quantity of the product, and a management team coordinating the overall operations. Yet, in the operations management community, the implication of *team decision making* has not been well understood. Researchers in behavioral operations management (BOM) have predominately used *individuals* to conduct experiments and generate managerial insights. The goal of this paper is to take a first step at understanding how teams make operational decisions. To the extent that teams are found to make better operational decisions than individuals, firms should consider encouraging the use of teams.

There is good reason to suspect that teams might reach different decisions than individuals. Research in behavioral economics shows that teams are better at making *tactical* and *strategic* decisions. Tactical decisions are those that arise in pure problem-solving tasks and does not involve interacting with other players. Teams are found to be more cognitively sophisticated than individuals and better at understanding the probabilistic nature of the decision tasks [Charness et al. (2007), Charness et al. (2010)]. Strategic decisions are those that involve considering other players' behavior when making your own decisions. Teams are found to have stronger self-interested preferences (more selfish) and are able to reason from the opponent's perspective. As a result, teams perform more consistently with game theoretic predictions [Cox (2002), Kocher and Sutter (2005), Cooper and Kagel (2005)]. While the above studies in behavioral economics have generated useful insights, they are all conducted using relatively abstract decision contexts. In our paper, we explore to what extent these insights carry over to an operations decision-making setting. To study how teams make operational decisions and how they compare to decisions made by individuals, we study settings with very well-established benchmarks for individual operational decision-making. This leads us to use Newsvendor decisions as our tactical decision context, and information sharing in a Newsvendor context as our strategic decision context. Both are important in practice and have been studied in BOM with individuals as subjects [Schweitzer and Cachon (2000), Özer et al. (2011)], providing a clear benchmark for individual behavior.

This paper is among the first in the BOM literature to study how teams make decisions, and we need to choose which *team setup* to use. In this paper we consider two-member teams with aligned information, aligned preferences, and consensus-based decision making. We allow team members to discuss freely with each other when making decisions.[1] This setup has the following advantages. (1) A small team setup (of two team members) is simple to work with, and past research has shown that increasing the team size from team of one to two is enough to generate significant behavioral differences [Cooper and Kagel (2005), Charness et al. (2007)]. (2) This setup allows us to focus on the question of whether or not teams, when fully aligned, make better decisions than individuals, and removes the effect that teams may perform worse than individuals due to misaligned incentives within the team, or untruthful sharing of information.[2] (3) The consensus-based decision making scheme encourages team members to reflect their reasoning processes in their chats. By linking their chats with the team's decision outcome, we are able to directly study the *decision drivers*. Studying the underlying decision drivers has proven to be very challenging with individual subjects [Thonemann and Becker-Peth (2018)].

[1] Team discussions occur via text chat, and are recorded by the experimental software. Analyzing the discussions will be a key part of our analysis.

[2] Also, based on our conversations with the large consumer goods firm, in practice much effort is devoted to align team members' information and incentives, in the hopes of reaching better outcomes. Thus, our study, which presupposes alignment and common information, can be construed as giving the teams the best possible shot to perform well.

In the standalone Newsvendor decision we hypothesize that, compared to individuals, teams will make better Newsvendor decisions; surprisingly, we find that teams make *worse* Newsvendor decisions compared to individuals. Past research has shown that individuals typically fail to make the Newsvendor decisions optimally: they often order too close to the mean and are hence subject to the "pull-to-center" bias [Schweitzer and Cachon (2000), Thonemann and Becker-Peth (2018)]. We find that teams are more strongly affected by the "pull-to-center" bias. For the Newsvendor under information sharing, we employ the setting proposed by Özer et al. (2011): a retailer (he) privately observes the demand forecast and strategically sends a demand signal to the supplier (she); the supplier updates her belief regarding the demand and makes the Newsvendor decision. The game theoretic prediction is that the retailer should always inflate to the highest possible amount when reporting the signals, regardless of the private demand forecast he receives; anticipating this, the supplier should never believe the signal. Interestingly, experiments have shown that individual retailers send out informative signals, and individual suppliers rely on the signals to update their beliefs [Özer et al. (2011)]. In this paper, we hypothesize that teams are more untrustworthy and more untrusting compared to individuals, because teams tend to be more selfish. We find that team retailers are more likely to lie by "inflating more" when reporting the signals; hence, team retailers are more untrustworthy and behave more consistently with what game theory predicts. However, team suppliers and individual suppliers are equally trusting towards the information shared by the retailers.

An important research objective in behavioral operations is understanding the underlying behavioral processes that drive operational decisions; if these can be identified, they can be better managed. Our analysis with team chats allows us to externalize their decision-making processes in a way not possible in past work that studied individuals. We find that teams frequently refer to the *mean of demand* when making the Newsvendor decision, consistent with what is assumed in a widely discussed Newsvendor decision theory: *Mean Anchoring and Insufficient Adjustment Towards the Optimum* [Schweitzer and Cachon (2000)]. Yet, our analysis suggests that teams do not hold a firm belief regarding the optimal direction to adjust to (higher/lower quantity), and instead their decisions are driven by conversations emphasizing a desire to be aggressive or conservative in that round. We also find that while many teams put effort into formulating concrete Newsvendor decision strategies, doing so does not improve either the quality or consistency of their decisions. In the Newsvendor with information sharing, we find that: on the retailer side, the argument of being untrustworthy dominates the argument of being trustworthy; on the supplier side, both the argument of being trusting and untrusting are impactful, and their effects cancel out at an aggregate level. These findings from chat analyses help us explain the decision outcomes we observed in our experiments.

The notion of a “Eureka!” type decision helps us organize these findings. The Eureka-type decision has two key characteristics: (1) there is a clear, unambiguous optimal solution to the question/decision task; (2) subjects within the same team are able to easily demonstrate or verify the optimality of a proposed solution, which allows them to efficiently converge to the optimal solution. This type of decision has an “Aha!” moment. Typical examples include insight problems[3] such as word puzzles, and the strategic decision in the classical prisoner’s dilemma with a dominant strategy to confess. Past research in social psychology and behavioral economics suggests that if the decision is Eureka-type, teams will perform better than individuals for the reason that once one member of the team gets the correct answer, he/she can easily explain the reasoning process to his/her teammates and convince them [Davis (1992), Cooper and Kagel (2005)].

In the context of our paper, the decision tasks subjects faced share some key components of the Eureka-type decision, but they also differ in important ways, making it difficult to generate hypotheses based on this concept *ex-ante*. Chat analysis helps us gain better understanding into teams’ thought processes, which provides direct evidence *ex-post* regarding whether the decision is Eureka-type. In the standalone Newsvendor decision, the optimal solution exists; however, it is unclear whether the conceptual framework of the optimal solution to appropriately balance overage and underage costs will be apparent and salient to subjects. Chat analysis shows that teams do not benefit from their effort to formulate Newsvendor decision strategies, and they are unable to converge to the optimal solution through discussions. Hence, the standalone Newsvendor decision is non-Eureka, and we find that teams perform worse than individuals. In the Newsvendor under information sharing, game theory gives a clear equilibrium prediction, but derivation of the retailer’s and the supplier’s optimal strategies requires different levels of cognitive sophistication. Chat analysis shows that team retailers are able to demonstrate and verify the the simple strategy to “inflate from X”, while team suppliers have a hard time formulating their optimal responses. Hence, the retailer’s decision is Eureka-type, and we observe that team retailers are more strategic; the supplier’s decision is non-Eureka, and we find that team suppliers and individual suppliers perform similarly. In general, the “Eureka-type decision” concept is useful in terms of organizing and understanding our findings; but, as we have discussed above, in many managerial and operational contexts it may be difficult to identify *ex-ante* whether a decision is Eureka-type, and chat analysis proves to be a useful tool to make this identification *ex-post*. We provide a more comprehensive discussion in Section 7.

[3] Insights usually involve “the sudden emergence of a solution into awareness as a whole — a ‘great speculative leap’ — in which the processes leading to solution are unconscious and can be consciously reconstructed only after the fact” [Salvi et al. (2016)].

The remainder of this paper is arranged as follows: Section 2 reviews related literature. Section 3 sets up the decision contexts and hypotheses. Section 4 explains the experiment design. Section 5 presents the experiment results. We analyze the team decision mechanism using chat analysis in Section 6. Section 7 discusses the main results. Section 8 concludes the paper.

2. Literature Review

In the past twenty years, the behavioral operations management (BOM) has developed into an established field that studies how human subjects make managerial decisions in various operational contexts. Well-known research includes bias in the forecasting decisions [Kremer et al. (2011), Kremer et al. (2015)], Newsvendor decisions [Schweitzer and Cachon (2000), Ho et al. (2010)], contracting and procurement [Katok and Wu (2009), Wan et al. (2012)], and trust in the supply chains [Özer et al. (2011), Özer et al. (2014), Beer et al. (2017)]. For an excellent summary of the current state of BOM, please refer to the forthcoming handbook in behavioral operations management [Donohue et al. (2018)]. To date, BOM researchers have predominately used individuals, not teams, as subjects to run the experiments. It is unclear to what extent these managerial insights will hold in practice where many important managerial and operational decisions are made by teams.

Research in behavioral economics suggests that teams do behave differently from individuals in systematic ways. Researchers have studied team decision making mainly in two contexts: tactical decision making and self-interested strategic decision making. The team setup we use in our experiments has been the primary format for past tactical/strategic experiments: small teams, aligned information, aligned preferences, and consensus-based decision making scheme; hence, we expect that the insights from these studies should extend to operational contexts. For tactical decisions, teams are found to be more *rational* in the sense that they are more cognitively sophisticated and make fewer errors. Charness et al. (2007) consider a decision task where subjects learn the state of the world by drawing balls. They find that teams are better at performing Bayesian updating and making decisions based on the principle of first-order stochastic dominance. Charness et al. (2010) study teams' performance in the "Linda paradox" proposed by Tversky and Kahneman (1983). Subjects are given two options that describe the characteristics of lady Linda, and they are asked to pick the more probable one. Option A has one fewer constraint than option B and they are otherwise identical. From a statistical point of view, option B is less probable; yet, if subjects are affected by the conjunction fallacy, they may end up choosing option B. Charness et al. (2010) find that while individuals predominately choose option B, teams are much more likely to choose option A. Interestingly, in our paper we find that teams' advantage in making tactical decisions does not extend to the operational context of making Newsvendor decisions. We find that teams are *more* *affected* by the pull-to-center bias compared to individuals. That is, teams are more likely to make an inventory quantity decision that is too close to the mean of demand and is further away from the optimal Newsvendor quantity.

For strategic decisions, several studies have shown that, compared to individuals, teams are found to be more *selfish* and *strategic*; teams pay more attention to monetary payoffs, and act much more consistently with what game theory predicts [Charness and Sutter (2012)]. In the market entrant signaling game where the weak incumbent can pretend to be the strong incumbent by setting a high production quantity, Cooper and Kagel (2005) find that teams are much more likely to play the equilibrium strategy, and this is driven by their ability to reason from the opponent's perspective. In the beauty-contest game where subjects are asked to picked a number that is p times the average of all other subjects' numbers ($p < 1$), game theory predicts that everyone should pick the smallest possible number. Individuals typically pick a number that is far above the minimum, while teams tend to pick a significantly lower number [Kocher and Sutter (2005)]. In the trust game where the sender sends money to the receiver, followed by money returned by the receiver, a rational sender will send zero to the receiver because a rational receiver will not return any positive amount back to the sender. Hence, any deviation from the rational behavior must be due to the fact that the sender trusts the receiver and the receiver is being trustworthy. Individuals have been found to send and return significantly positive amounts of money, while teams tend to send or return lower amounts, showing a lower degree of trust or trustworthiness [Cox (2002), Kugler et al. (2007)]. In this paper, we study the trust and trustworthy behavior of teams and individuals in a supply chain context: the Newsvendor under information sharing. We find that teams are *less trustworthy* than individuals but *equally trusting.*

In behavioral Newsvendor studies a very robust finding, first identified in the seminal paper by Schweitzer and Cachon (2000), is that individuals fail to make the Newsvendor decision optimally. They often order too close to the mean, hence being subject to the "pull-to-center" bias.[4] Training, feedback and even years of working experience do not solve this problem [Bolton and Katok (2008), Bolton et al. (2012)]. Despite the fact that this observation is so robust, the driver behind the pull-to-center bias is still unclear. Potential explanations include mean anchoring and insufficient adjustment [Schweitzer and Cachon (2000), Bostian et al. (2008)], psychological costs for waste and stock-out in inventory [Schweitzer and Cachon (2000), Ho et al. (2010)], impulse balance [Ockenfels and Selten (2014)], and bounded rationality [Su (2008)]. With chat analysis, we are able to study whether the proposed explanations are consistent with the reasoning processes in teams. In Newsvendor experiments, identifying the actual underlying thought process has proven to be very

[4] For a list of dozen plus papers confirming this observation, please refer to Figure 1 in the review chapter by Thonemann and Becker-Peth (2018).

challenging [Thonemann and Becker-Peth (2018)]. The team chats we collect help to externalize teams' reasoning processes and draw clear conclusions. Of course, it is possible that teams and individuals are using different reasoning processes; but, given that *both* teams and individuals are subject to the pull-to-center bias, there are reasons to believe that they share similarities in their reasoning processes.

For the Newsvendor under information sharing, Özer et al. (2011) show that trust and trustworthiness play a critical role in maintaining effective information sharing, and Özer et al. (2014) extend this insight to a cross-cultural setting. In our paper, we utilize a similar setup as Özer et al. (2011), but adjusted for teams making decisions. This allows us to utilize the insights of Özer et al. (2011) as a strategic decision making benchmark for individuals, against which we can compare our results for teams. The contrast between individual and team decision-making is explored throughout our paper, and unlike Özer et al.'s paper, the usage of teams also affords us the opportunity to externalize the decision-making process.

Given its importance, the topic of team decision making is relatively under-studied in BOM. In our study we find that teams making Newsvendor decisions perform worse than individuals, and are less trustworthy but equally trusting in the information sharing game. Of course, there is scope for additional studies. To our knowledge there are just three existing papers that examine team decision making in operations, in a Newsvendor setting, albeit under various experimental interventions: providing the subjects with multiple decision proposals [Gavirneni and Xia (2009)], endowing one team member with superior information [Laya and Pavlov (2015)], or placing subjects in a multi-round tournament setting where contestants (teams or individuals) are in competition and learn the winning decision made each round [Wu and Seidmann (2015)]. Our focus is different — we study a standard Newsvendor setting, along with the information sharing game. To align with the information sharing game, our Newsvendor setting is distinct in that the mean of demand changes in each round. We also design our study to uncover the decision mechanisms underlying the observed decisions and match them against existing behavioral models of decision making, using the team chats. Interestingly, as we discuss briefly in Footnote 23, our findings are able to help organize some of the observations of team performance in these three papers, which find that team Newsvendors largely perform similarly to individuals, save one (of four) treatments in Wu and Seidmann (2015) where teams perform slightly better.

3. Decision Contexts and Hypotheses

In this section, we introduce the decision context of our paper, which is a modified version of the information sharing game proposed by Özer et al. (2011). As we will explain in this section, the information sharing game, when appropriately adjusted for experimental proposes (as in Özer et al.

(2011)), allows us to study both the standalone Newsvendor decision and the Newsvendor under information sharing.

There is one supplier (she) and one retailer (he) in this information sharing game. The supplier produces for the retailer who sells to the end customers. The end customer demand is uncertain. Relative to the supplier, the retailer has better demand information due to his proximity to the market. The retailer determines how truthfully he wants to share his demand information with the supplier. The supplier interprets the information shared by the retailer and make her production decision. This is a one-shot game, so reputation does not play a role. To ensure this in our experiments, retailers and suppliers are anonymous and they are randomly re-matched from round to round.

Formally, the end customer demand D equals $X+\xi$, where X and ξ are random variables. X is known as the *demand forecast*, and it follows the cumulative distribution function (c.d.f.) $F(\cdot)$ over support $[X_l, X_u]$. ξ is the market uncertainty and is distributed on $[\xi_l, \xi_u]$ with mean 0 and c.d.f. $G(\cdot)$. We assume $X_l + \xi_l > 0$ to ensure a positive demand. In our experiments, X is uniformly distributed between 100 and 400; ξ is uniformly distributed between -75 and +75.

The event sequence of the game is as follows: (1) At the beginning of the game, the retailer privately observes the realized value of X, and he delivers a signal $\tilde{X}$ to the supplier. The demand uncertainty ξ remains uncertain to both the retailer and supplier. (2) For the supplier, after she receives the demand signal $\tilde{X}$, she updates her belief about the end demand and makes her production decision Q and incurs a cost $c\cdot Q$. (3) The demand uncertainty ξ is realized. The retailer orders D from the supplier, receives $\min(D,Q)$ from the supplier, pays the supplier $w\cdot\min(D,Q)$, and sells these units to the end customers to obtain revenue $p\cdot\min(D,Q)$. To keep the context simple, all the price and cost parameters are exogenous. The specific parameters we use in the experiment are: $p=140, w=100, c=80$. The (prior) distributions of X and ξ, the event sequence, and the price/cost parameters are all common knowledge. The only piece of asymmetric information is the realized value of X privately observed by the retailer.

In this information sharing game, on the retailer side, notice that he bears no inventory risk and will not be punished for misreporting the signals. So, game theory suggests that it is always optimal for him to report X_u, which is the highest amount possible, regardless of the private demand forecast X he receives. This also means that for the retailer there is a clear and simple strategy to "inflate from X" when reporting the signal $\tilde{X}$. On the supplier side, anticipating this, she should never update her belief based on the signal $\tilde{X}$ she receives when making the production decision.[5]

In our experiments, all the subjects play the above information sharing game *in two different settings*. In the first setting, the "computerized retailer setting", the computer plays the retailer

[5] See Özer et al. (2011) for a formal proof of these results.

role and all the subjects (teams and individuals) play as the supplier. The computer always reports the signal *truthfully* to the supplier, and the supplier knows this. Hence, the supplier knows the exact demand distribution when making the production decision Q. This setting corresponds to the standalone Newsvendor setting, and it allows us to study how teams and individuals make the Newsvendor decision. In the second setting, the "human retailer setting", both the retailer and the supplier are played by human subjects. This setting corresponds to the Newsvendor under demand information sharing, and it allows us to study how teams and individuals make strategic decisions in the context of inventory planning. In fact, for each team/individual, we need to combine the data from *both* settings to determine how strategic the team/individual is. This point will be made clear in Section 3.2. We now present the details of each setting.

3.1. The Computerized Retailer Setting

In the computerized retailer setting, because the computerized retailer always reports the signal truthfully and the supplier knows this, it is straightforward for the supplier to update her belief regarding the customer demand: The mean of demand is just the signal she receives. Let Q_{CR} be the supplier's production decision in the computerized retailer setting. Notice that X changes from round to round while the interval of demand uncertainty ξ does not. Namely, the optimal decision can be expressed as a function of X. With the parameters introduced above, the optimal production decision Q^*_{CR} is: $Q^*_{CR} = X - 45$. Hence, we can use $(X - Q_{CR})$, the production adjustment relative to the mean X, to measure the supplier's performance in making the standalone Newsvendor decision in a given round. In the experiment, subjects make the standalone Newsvendor decision multiple times; so, the average value $\overline{(X - Q_{CR})}$ measures the supplier's overall ability in making the Newsvendor decision.

We now hypothesize teams' behavior for the Newsvendor decision. Past literature has consistently shown that individuals fail to make the Newsvendor decisions optimally. Individuals often order too close to the mean, and hence are subject to the so-called pull-to-center bias [Schweitzer and Cachon (2000)]. Given that teams are often better at making tactical decisions, we conjecture that they will also make better Newsvendor decisions. We propose the following hypothesis:

Hypothesis 1. *With the computerized retailer, teams perform better in making Newsvendor decisions compared to individuals, i.e., teams' production decision Q_{CR} will be closer to the optimal production decision Q^*_{CR}.*

3.2. The Human Retailer Setting

In the human retailer setting, both the retailer and the supplier are played by human subjects. Let $\tilde{X}$ be the signal sent by the retailer, and Q_{HR} be the production decision made by the supplier in the human retailer setting after receiving $\tilde{X}$.

On the retailer's side, it is always optimal for the retailer to report highest amount possible X_u, regardless of the true value of X. However, past behavioral research suggests that people care about trust relationships between each other, and this is true even in non-repeated game settings among strangers or anonymous participants [Fehr et al. (1998), Charness et al. (2004)]. Hence, the retailer's decision is determined by how *trustworthy* he wants to be. A fully *trustworthy* retailer will simply set $\tilde{X} = X$ to tell the truth to the supplier. An *untrustworthy* retailer will distort the signal to act in his own benefit: The more untrustworthy he is, the more he will inflate the signal towards X_u. We use the degree of inflation $(\tilde{X} - X)$ to measure his degree of (un)trustworthiness.

On the supplier's side, she needs to consider two questions in order to determine Q_{HR}: (1) to what extent she trusts the signal sent by the retailer, i.e., how to update her belief regarding the end customer demand; and (2) how to make the optimal production decision with the updated demand distribution. Her response to the first question depends on how trusting she is. If she is fully *trusting* towards the retailer, she will believe the signal to be true and set it to be the updated mean of demand. If she is fully *untrusting*, the signal is useless and she does not update her belief. Her response to the second question depends on how sophisticated she is in making the standalone Newsvendor decision.

In the human retailer setting, we only observe the supplier's final production decision Q_{HR}, which reflects her answer to both of the questions jointly. To disentangle the supplier's degree of *trusting*, we need to contrast her production adjustment $(\tilde{X} - Q_{HR})$ in the human retailer setting with her average production adjustment in the computerized retailer setting $(\overline{X - Q_{CR}})$, similar to Özer et al. (2011). Then, the difference $(\tilde{X} - Q_{HR}) - (\overline{X - Q_{CR}})$ measures the supplier's trusting behavior. If the supplier is very trusting towards the human retailers, her behavior in the two decision contexts should be quite similar, indicating that $(\tilde{X} - Q_{HR}) - (\overline{X - Q_{CR}})$ should be close to 0. On the other hand, the more *untrusting* she is, the more she thinks the signal is inflated and adjusts further away from it by setting a lower Q_{HR}, which translates into a higher value of $(\tilde{X} - Q_{HR}) - (\overline{X - Q_{CR}})$.

The behavioral economics literature shows that teams, broadly speaking, have stronger self-interested preferences and are more untrustworthy/untrusting towards the opponent [Charness and Sutter (2012)]. Translating into our context, this suggests that in the human retailer setting, teams should be more untrustworthy as the retailer and more untrusting as the supplier. But, because teams also tend to show a certain degree of trust and trustworthiness [Cox (2002), Kugler et al. (2007)], we conjecture that they will not go so far as either to set the signal to be uninformative as the retailer or completely disregard the signal as the supplier. Finally, in the human-retailer setting, the identity of the opponent (team or individual) may also affect the degree of trust and trustworthiness. In other words, a supplier (retailer) may behave differently depending upon whether the retailer (supplier) she faces is being played by an individual or a team. Because past research shows mixed results in this part [Cox (2002)], we make the conservative conjecture that the above intuition holds regardless of the identity of the opponent. We propose the following two hypotheses:

Hypothesis 2. *In the human retailer setting,*

(a) The signal $\tilde{X}$ sent by the retailer is positively correlated with the private demand forecast X he observes.

(b) The supplier's production decision Q_{HR} is positively correlated with the signal $\tilde{X}$ she receives from the retailer.

Hypothesis 3. *In the human retailer setting,*

(a) As the retailer, regardless of the identity of the opponent (team or individual), teams tend be more untrustworthy compared to individuals. In other words, $(\tilde{X} - X)$ is, in general, higher for team retailers than for individual retailers.

(b) As the supplier, regardless of the identity of the opponent (team or individual), teams tend be more untrusting compared to individuals. In other words, $(\tilde{X} - Q_{HR}) - (\overline{X - Q_{CR}})$ is, in general, higher for team suppliers than for individual suppliers.

To summarize, the computerized retailer setting measures the behavior of teams and individuals in the standalone Newsvendor decision; the human retailer setting studies the Newsvendor under information sharing. The two decision contexts *together* allows us to measure the degree of trust and trustworthiness of teams and individuals.

4. Experiment Design

As we have shown in Section 3, the information sharing game allows us to study the Newsvendor decisions and trust decisions by running experiments with the two decision contexts: computerized retailer and human retailer. Hence, in the experiment, we will set up a treatment where teams make decisions in these two contexts, and compare it with a benchmark treatment where individuals make decisions in these two contexts. Comparing teams' and individuals' performances with computerized retailers allows us to test Hypothesis 1. Comparing teams' and individuals' performances with human retailers allows us to partially study the trust and trustworthiness behavior of teams, which corresponds to Hypotheses 2 and 3. To fully test the Hypotheses 2 and 3, we also need to control the *opponent identity* (team or individual). Hence, we set up the third treatment - the mixed treatment, wherein we have teams playing against individuals.

The above design with 2 decision contexts and 3 treatments is a complete design that allows us to test all the hypotheses we have raised. Finally, we set up an additional decision context, which we call “training” at the beginning of the experiment. In this context, subjects play 3 rounds of the computerized retailer setting *individually*, regardless of the treatment they are in. This training serves two purposes: (1) gets subjects familiar with the context at the beginning of the experiment; (2) collects subjects’ individual behavior and preferences in making the Newsvendor decisions with a computerized retailer. This data allows us to study how individuals’ preferences are integrated within a team when making the standalone Newsvendor decision.

Figure 1 Experimental Design

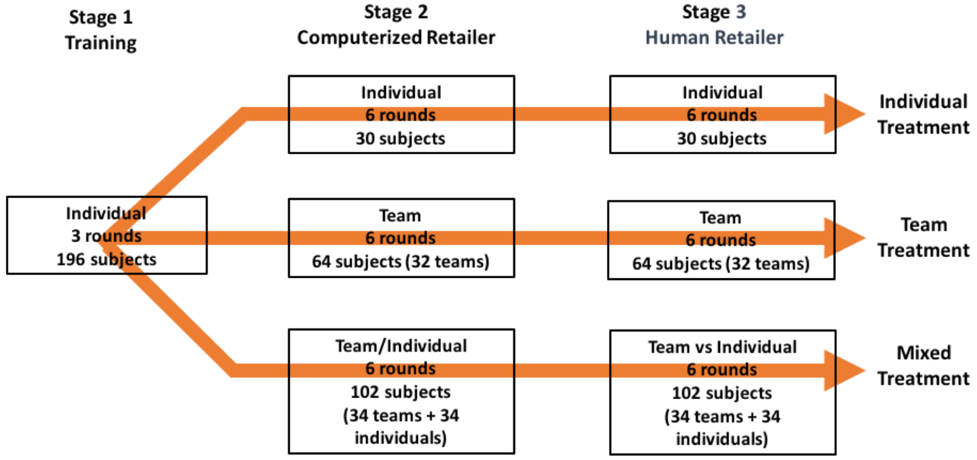

Figure 1 summarizes the full experimental design. Each of our 3 treatments share the same timeline. The timeline is divided into three stages: (1) In stage 1, subjects play 3 rounds of the computerized retailer setting as individuals, regardless of the treatment they are in. (2) In stage 2, subjects continue to play 6 more rounds of the computerized retailer setting in either teams or as individuals, depending on the treatment they are in. (3) In stage 3, subjects play 6 rounds of the human retailer setting. In the human retailer setting, the roles (supplier or retailer) are randomly assigned at the beginning of each round. To control the reputation effect, the matching between the retailer and supplier is anonymous. Consecutive pairings are disallowed; but otherwise, the matching is random in each round. The 3 treatments we developed are given below:

1. **The individual treatment:** Subjects make decisions individually throughout the experiment. This can be considered a replication of the experiment in Özer et al. (2011).

2. **The team treatment:** Subjects are formed into two-person teams at the beginning of stage 2, and they keep making decisions in that team for the remainder of the experiment. So, in this treatment in the human retailer setting, we will have teams playing against teams.

3. **The mixed treatment:** At the beginning of stage 2, two-thirds of the subjects form two-person teams. The remaining one-third continue decisions as individuals for the remainder of the experiment. This assignment stays unchanged for the remainder of the experiment. In this

treatment in the human retailer setting, we always have either team retailers playing against individual suppliers or the other way around.

The experiment timeline and treatment are public information for all the subjects. That is, subjects know whether they will play as a team/individual and whether they are playing against a team/individual. At the end of each round, subjects receive the following feedback information in a dashboard: the decision they have made in that round, the realized end customer demand, the profits they have earned in that round and their accumulated profit over the previous rounds. To assist comparisons across treatments, in the experiment, we pre-generate a list of X and ξ, and apply them to all the three treatments. For example, in round 1 of the computerized retailer setting, all suppliers in the three treatments observe the same value of X at the beginning of that round and observe the same realized demand at the end of the round. The experiment is conducted in an economics laboratory in a large public university in the United States. The participants are undergraduate and graduate students from the university. The participation time is around 1 hours and 15 minutes. Subjects gain 5 dollars show-up fee upon joining the experiment and earn additional payments based on their performance in the experiment. The average payoff is 20 dollars. We use Z-tree to program the experiment [Fischbacher (2007)] and ORSEE for subject recruitment [Greiner (2004)].

4.1. The Setup of Teams

In this subsection, we give a detailed description of how teams are formed and how the team members interact to make decisions. In the team treatment and mixed treatment, at the beginning of stage 2 (computerized retailer setting), the subjects are randomly paired to form two-person teams. They keep making decisions in the same team for the remainder of the experiment. The information is *fully transparent inside a team*, meaning that the two team members observe the same information and get the same feedback. To keep the monetary payoff comparable across treatments, we set each team member's payoff the same as the team's payoff, i.e., the payoff is not split between the two team members. Within each team, the two members interact with each other using a computer interface. To facilitate team decision making, the two members are allowed to text chat with each other.[6] This is a way to help us externalize their thought processes. In Section 6, we will make use of these chats to study teams' decision mechanisms.

A key feature of team decision making is how the decisions are reached. In our experiments, we employ a consensus-based decision making scheme: The two team members need to jointly agree on a value (Q if supplier, $\tilde{X}$ if retailer) in order to form a team decision. The detailed rule is as follows: The team decision procedure consists of two types of actions — making a proposal

[6] There is no cross-team chat communication.

and agreeing to a proposal. When a round starts, the two team members chat with each other to discuss, and they can make numerical proposals to reflect their thoughts. The latest proposals from both of them will be shown on the screen. The final team decision is made when one team member picks the latest proposal from the other team member and confirms it. If no confirmation is made within the given time limit, the computer will randomly pick one member from the team and use his/her latest proposal as the team's decision. If the round ends before either team member makes a proposal, the computer will randomly generate a number between 25 and 475. Both teams and individuals are given two minutes in each round. This is to give teams enough time to discuss and reach mutual agreement, while also keeping the time allotted for decisions identical across all 3 treatments. Subjects are given one additional minute in the *first round* in stage 2 and stage 3. Each round ends once all the decisions have been made. In our experiments, on average teams make 1.57 proposals in each round; more than 97% of teams are able to reach an agreement within the given time limit. As for the time teams spend to reach an agreement, in both the standalone Newsvendor decision task and the Newsvendor under information sharing, on average teams spend less than 35%-45% of the time given. Hence, most team decisions come from collaboration and relatively quick agreement by members, rather than extensive internal conflict.

The consensus-based decision making scheme, along with aligned information and aligned preferences, has the advantages that it is simple to employ and it naturally represents many real-world team decision-making settings. Moreover, this design encourages team members to fully elaborate the reasoning behind each proposal using text chats in order to reach a final team decision. Hence, the team chats provide us with the information that reflects the team members' thought processes, which we can directly relate to the observed decisions reached by teams.[7]

5. Experiment Results - Decision Outcomes

In this section, we present the results on decision outcomes. The discussion of chat analysis will be presented in Section 6. We first present the results for the computerized retailer setting.

5.1. Computerized Retailer Setting

Recall that the optimal ordering decision is $Q^*_{CR} = X - 45$; thus, the pull-to-center bias means that subjects order close to the mean X and do not make a large enough reduction from the mean[8]. If Hypothesis 1 is true, teams should perform better by making *larger* reduction from the mean and being *closer* to the optimum.

[7] Gavirneni and Xia (2009) also study subjects' decision processes, but does so by asking them to justify their decisions *after* each decision has been made. Our design has the advantage that extensive team discussions happen *before* the decision is made, and such discussions are an integrated part of the decision-making process.

[8] Hereafter we simply use "reduction" to refer to "reduction from the mean" when appropriate.

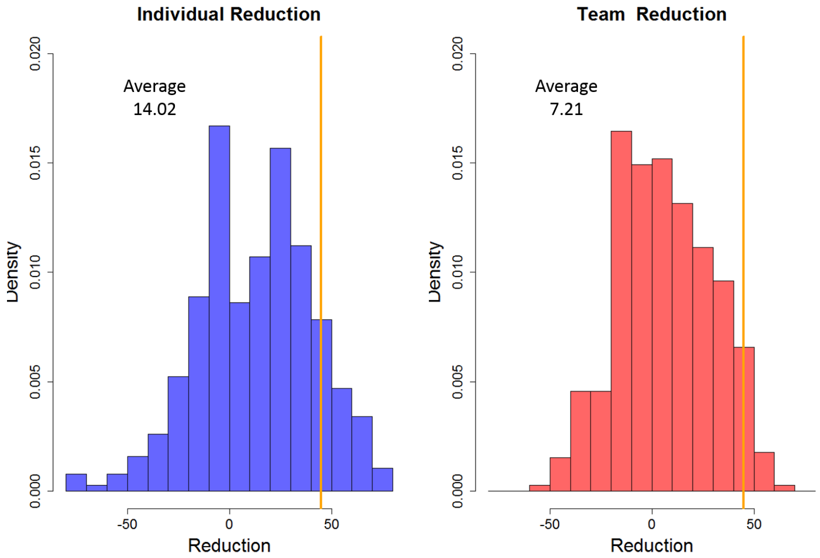

Figure 2 **Reduction in the Computerized Retailer Setting**

We present the results in Figure 2, which depicts the observed stage 2 reduction relative to the demand mean X for individuals and teams. The optimal level of reduction (45 units) is depicted using the orange lines. For individuals, we see that many of them make no reduction from the mean, and there are also many individuals choosing reduction quite close to the optimum. On average, individuals reduce 14 units from the mean, which is smaller than the optimum and is too close to the mean, consistent with the pull-to-center bias. For teams, quite surprisingly, there is even more pull-to-center bias exhibited, and on average teams reduce by 7.54 units, which is half as much as individuals' average reduction. The Wilcoxon rank-sum test on average reduction confirms the difference to be significant ($p = 0.02$). To study this more formally, we run a random-effects regression analysis. The function we consider is given below:

$$Q_{CR,it} = \text{Intercept} + \beta_X \cdot X_t + \beta_{team} \cdot Team_i + \beta_t \cdot t + \epsilon_{it}. \tag{1}$$

In this and all the subsequent regressions, t refers to round t and i refers to the "decision-making unit" i. If subjects make decisions individually, i naturally refers to that subject. If subjects make decisions in a team, then i refers to the team. Given that two members inside a team share the same information to reach mutual agreement to produce a team decision and receive the same payoff, in the regression we only have *one* data point for each team in a round. The variable $Q_{CR,it}$ represents the decision-making unit i's production decision made in round t in the computerized retailer setting. X_t represents the demand forecast, which is also the mean of customer demand subjects receive in round t. Notice that we control for X_t to be identical for all teams/individuals in the same round. $Team_i$ is the dummy variable for the decision-making unit i being a team or not. The round variable t controls for time trends.

Table 1 Computerized Retailer Setting: Newsvendor Decision Regression Analysis

Variable	Estimate of the Coefficient
X_t	$1.00(0.01)^{***}$
$Team_i$	$6.81(2.96)^{**}$
t	$-0.20(0.56)$
Intercept	$-13.62(3.88)^{***}$

Notes: Random-effects GLS regression with balanced panel data, clustering on the decision unit level. 780 observations over 6 rounds, from 66 teams and 64 individuals. Robust standard errors are in parentheses. Significance is denoted: $^{*}p < 0.1,^{**} p < 0.05$, $^{***}p < 0.01$.

Using regression (1) we employ a random-effects regression, clustering on the level of the decision-making unit. The results are summarized in Table 1. We see that the intercept is significantly negative, meaning that on average, subjects do reduce from the mean to set their production decisions. However, the estimate for β_{team} is significantly positive, indicating that teams on average reduce less from the mean, hence end up being further away from the optimum. With these results, we reject Hypothesis 1. Contrary to what we expected, teams actually make *worse* Newsvendor decisions compared to individuals.

5.2. What Makes teams Perform Worse in the Computerized Retailer Setting?

In this subsection, we study what drives teams to make worse Newsvendor decisions compared to individuals.[9] We first consider how preferences are integrated inside teams. Recall that subjects start by making 3 rounds of Newsvendor decisions *individually* in stage 1 (training) in all the three treatments. This has an important implication for analyzing team decisions since we can learn each subject's individual characteristics before they are formed into teams. Specifically, with the data from stage 1, we can: (1) identify the more capable person and less capable person (in making the Newsvendor decision) within each team;[10] (2) use each subject's average production adjustment in stage 1 to represent his/her *individual* inclination in making the Newsvendor decision. Then, by relating how these two members act individually in the training stage (stage 1) to how the team they form behaves in the computerized retailer setting (stage 2), we can study how preferences are integrated inside teams.

Three possibilities can lead to the outcome that teams perform worse than individuals: (1) The team decision outcomes are primarily driven by the *less* capable person inside the team; (2) the team decision outcomes are primarily driven by the *more* capable person inside the team, but

[9] Here we study this problem by analyzing the decision data only. An analysis based on the text chats will be provided in Section 6.

[10] We identify the more capable person as the one who makes larger reduction from the mean in the first three individual rounds. In our data subjects almost always make insufficient reduction from the mean, so larger reduction means better performance.

there is loss in optimality when the two members try to reach an agreement; (3) the team decision outcomes are unrelated to either of the team members, i.e., teams make random decisions. To study this formally, we consider the following linear regression:

$$\overline{ReductionStageTwo_i} = \text{Intercept} + \beta_1 \cdot \overline{ReductionStageOneLC_i} + \beta_2 \cdot \overline{ReductionStageOneMC_i} + \epsilon_i. \tag{2}$$

The variable $\overline{ReductionStageTwo_i}$ refers to the average reduction for team i in stage 2. $\overline{ReductionStageOneLC_i}$ refers to the average reduction in the three individual rounds in stage 1 (training) for the *less capable person* of the team. $\overline{ReductionStageOneMC_i}$ refers to the average reduction in the three individual rounds in stage 1 for the *more capable person* of the team. β_1 and β_2 measure their relative weights on the team's decision outcome in stage 2.

We find that the estimates for neither the Intercept nor β_1 is significant. The estimate for β_2 is significant ($p = 0.04$), with the value of 0.28. Hence, we find evidence that the team decision outcome is mainly driven by the *more capable person* of the team, which is desirable. However, the more capable person is not able to fully turn his decision into the team's decision. On the other hand, the less capable person pulls the team's decision towards the mean X, by an amount independent of his/her own inclination (since $\hat{\beta}_1$ is insignificant).

Another aspect worth considering is the *learning dynamics*. Cooper and Kagel (2005) find that, in a market entrant game, teams converge to the equilibrium much faster than individuals. They attribute this to teams' ability to think from the opponent's perspective. Given the tactical nature of the Newsvendor decision, it is unclear whether teams will still exhibit this advantage.

From Table 1, we observe that there is not a statistically significant time trend for teams and individuals combined. This is also true when we run regression (1) for teams and individuals separately.[11] Hence, teams' production decisions do not become closer to the optimum over time. We also find that compared to individuals, teams are less willing to experiment with different decisions. We calculate the *variance* of the reduction from the mean over the 6 rounds in stage 2 for each decision unit, and compare the distribution of variances for teams and individuals. We find that teams have significantly smaller variances (Wilcoxon rank-sum test $p = 0.02$). This would not be a problem if teams are making better decisions. However, the data suggests that teams are more affected by the pull-to-center bias starting from the *first round* in stage 2 (Wilcoxon rank-sum test $p = 0.02$). Hence, the fact that teams are not willing to adjust makes the problem worse. In general, we find that teams do not have an advantage in terms of learning the optimum when making the standalone Newsvendor decision.

[11] P-value for the estimate of t is 0.70 and 0.89, respectively.

5.3. The Human Retailer Setting

We now proceed to the human retailer setting. In this subsection we will focus on analyzing the identity effect (team v.s. individual) of the decision maker. The opponent effect does not change the overall results and therefore is relegated to Appendix C. Recall that if Hypothesis 2 is true, the retailer's signal $\tilde{X}$ should be positively correlated with the private forecast X they receive, and the supplier's production decision Q should be positively correlated with the signal $\tilde{X}$ they receive. If Hypothesis 3 is true, compared to individuals, team retailers should inflate more and the team suppliers should reduce more.

We begin by testing Hypotheses 2a and 3a, by focusing on the retailer's behavior. Figure 3 depicts the level of inflation (the signal relative to the private forecast) by the individual (left panel) and the team (right panel) retailers. For individuals, we observe that the majority of them make no inflation at all. That is, they simply report their signals truthfully by setting $\tilde{X} = X$. On the team side, we see that majority of teams inflate a positive amount, and the average inflation is higher than that of individuals, consistent with Hypothesis 3a.

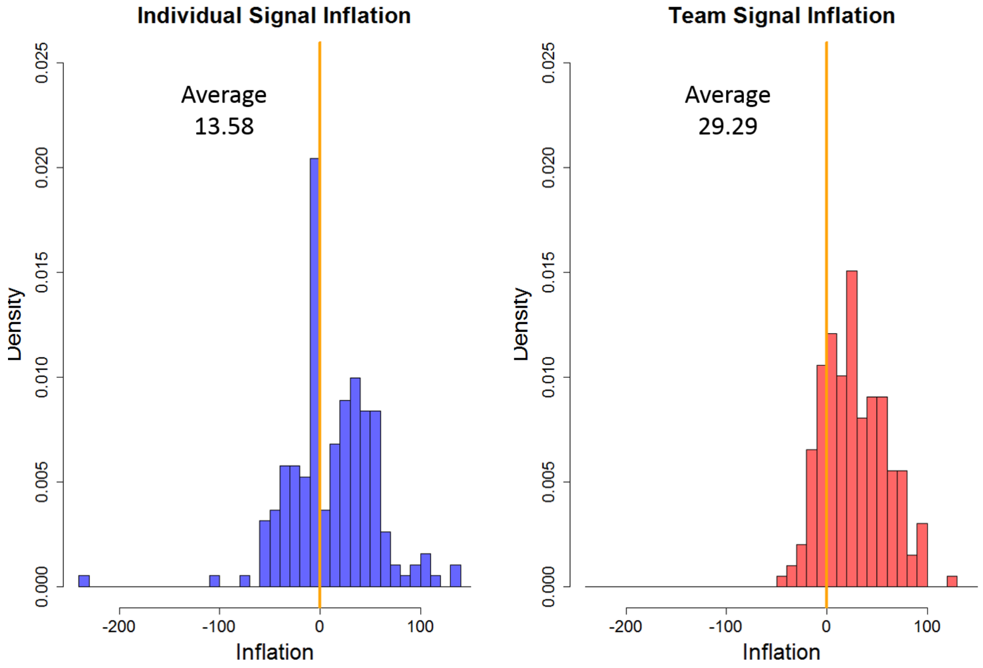

Figure 3 Human Retailer Setting: Retailer Inflation

To explore this more formally, we consider the following regression function:

$$\tilde{X}_{it} = \text{Intercept} + \beta_X \cdot X_t + \beta_{team} \cdot Team_i + \beta_t \cdot t + \epsilon_{it}. \tag{3}$$

The variable $\tilde{X}_{it}$ refers to the signal that the retailer i sends in round t. X_t refers to the private demand forecast the retailer receives in round t. $Team_i$ is the dummy variable for retailer i being a team or not. Variable t is used to capture the time trend. The regression results are summarized in Table 2.

Table 2 Human Retailer Setting: Retailer Signal Decision Regression Analysis

Variable	Estimate of the Coefficient
X_t	$0.88(0.03)^{***}$
$Team_i$	$15.13(5.71)^{***}$
t	$2.55(0.80)^{***}$
Intercept	$41.91(8.12)^{***}$

Notes: Random-effects GLS regression with unbalanced panel data. 390 observations over 6 rounds, from 66 teams and 64 individuals. Robust standard errors in parentheses. Significance is denoted: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

From the regression, we observe that the estimate for β_X is significantly positive and is close to 1, which confirms Hypothesis 2a. The estimate for the Intercept is significantly positive. These two results (informative and inflated signals) are consistent with the results in Özer et al. (2011) and Özer et al. (2014). On top of this, we find that team retailers inflate even more compared to individuals. This observation is also robust after we condition on the identity of the opponent. This confirms Hypothesis 3a, that team retailers do tend to inflate more, showing a tendency towards being less trustworthy compared to individuals. Finally, we observe a significantly positive time effect, which indicates that the retailers tend to inflate more over time.[12] In Appendix C, we show that the results are robust to the identity of the opponent; moreover, we conduct profit analysis and show that team retailers earn significantly higher profits compared to individual retailers: Applying random-effects GLS regression to profit comparisons, we find that, on average, team retailers earn 460 units more per round compared to individual retailers (p-value<0.01). That is, team retailers' (more) untrustworthy behavior translates into actual gains for them. Hence, team retailers have a decision advantage over individual retailers.

Now we turn to the supplier side to test Hypotheses 2b and 3b. First, to test Hypothesis 3b, we are interested in seeing whether team suppliers tend to be less trusting and hence reduce more from the *signal* they receive, after controlling for the heterogeneity in Newsvendor behavior. In Figure 4 we first present the supplier's degree of reduction relative to the signal they receive.

By comparing Figure 2 with Figure 4, we see that both the individual and team suppliers reduce more in the human retailer setting (stage 3) than in the computerized retailer setting (stage 2), indicating that both teams and individuals are able to respond to the inflation in the signal they receive from the human retailer. The Wilcoxon signed-rank test confirms the difference to be

[12] We test this time trend for teams and individuals separately and find it to be significant in both cases. Hence, both teams and individuals become more untrustworthy over time.

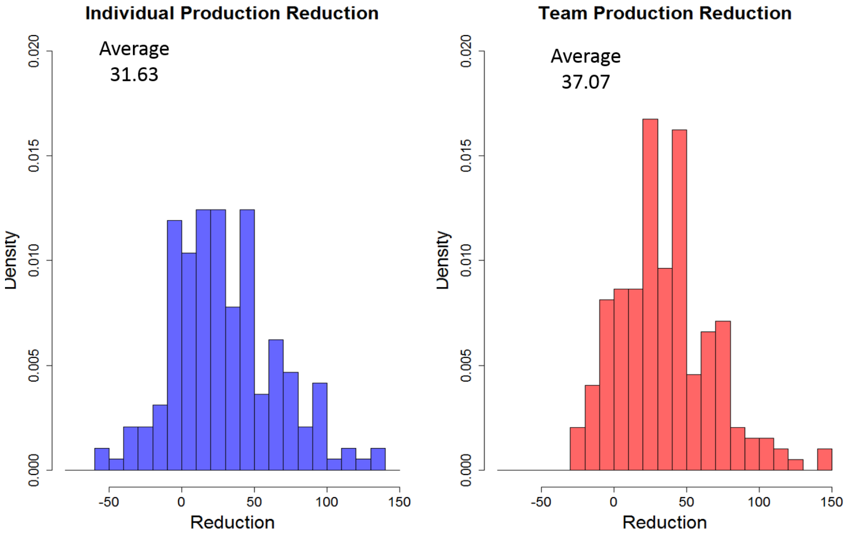

Figure 4 **Human Retailer Setting: Supplier Reduction**

significant for both teams and individuals ($p < 0.01$ in both cases).[13] However, in Figure 4 we do not see team suppliers behaving differently from individual suppliers, contrary to what we expected. To study the difference more formally, we consider the following regression function:

$$\begin{aligned} Q_{HR,it} =& \text{Intercept} + \beta_{signal} \cdot \tilde{X}_{it} + \beta_{team} \cdot Team_i \\ & + \beta_{NV} \cdot \overline{(X - Q_{CR})}_i + \beta_t \cdot t + \epsilon_{it}. \end{aligned} \tag{4}$$

The variable $Q_{HR,it}$ refers to supplier i's production decision in period t in the human retailer setting. $\tilde{X}_{it}$ refers to the signal that supplier i receives from the human retailer. We use $\overline{(X - Q_{CR})}_i$, supplier i's average Newsvendor reduction in the computerized retailer setting, to control for heterogeneity in Newsvendor behavior. The results are summarized in Table 3.

Table 3 **Human Retailer Setting: Supplier Production Decision Regression Analysis**

Variable	Estimate of the Coefficient
$\tilde{X}_{it}$	$0.89(0.20)^{***}$
$Team_i$	$-5.61(4.89)$
$\overline{(X - Q_{CR})}_i$	$-0.73(0.15)^{***}$
t	$0.31(0.73)$
Intercept	$8.31(7.17)$

Notes: Random-effects GLS regression with unbalanced panel data. 390 observations over 6 rounds, from 66 teams and 64 individuals. Robust standard errors in parentheses. Significance is denoted: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.

We observe that only two estimates are significant. The estimate for β_{Signal} is significantly positive and is close to 1, confirming Hypothesis 2b. β_{NV} is significantly negative, showing that the

[13] To meet the assumption of this test, we consider the *average* reduction in stage 2 and stage 3 decisions.

supplier's reduction in the human retailer setting is driven by her average Newsvendor reduction, which is not surprising. However, the estimate for β_{team} is not significant, showing that team suppliers do not reduce more from the signal compared to individual suppliers, contrary to what we expected. Hence, we find that team suppliers are *equally trusting* compared to individual suppliers, and we reject Hypothesis 3b. In Appendix C, we show that although the opponent's identity changes the supplier's degree of trusting, from a decision outcome's point of view team suppliers earn similar profits compared to individual suppliers: Applying random-effects GLS regression to profit comparisons, we find that the estimate of the coefficient associated with the $Team$ dummy variable is insignificant (p-value = 0.98). Hence, team suppliers do not have a decision advantage over individual suppliers.

To briefly summarize, we find that teams do behave differently from individuals in two aspects: Teams make worse Newsvendor decisions and they are more untrustworthy. The second finding is expected, because previous literature suggests that teams are more selfish. However, the first finding contrasts with previous literature that teams often make better tactical decisions. On the other hand, we find that teams are just as trusting as individuals. This is also contrary to what we expected, as past literature suggests that teams should be more skeptical towards the opponent. In order to explore the decision drivers of these findings, in the next section we make use of a novel approach: chat analysis.

6. Experimental Results - Decision Mechanism

Studying the decision mechanism has been the central topic in many behavioral studies; yet, it has proven to be challenging with individual subjects because the thought process cannot be directly observed. Researchers have tried different methods and found them either prohibitively expensive (e.g., the neuroscience approach) or inaccurate (for example the "think aloud" approach that requires people to write what they think when making decisions). In this section, with the help of team chats we are able to directly study the decision mechanism in a Behavioral OM problem. Recall that in our design we require team members to reach *mutual agreement* to form a team decision with the help of textual communication (chats). This design makes team chats an important part of teams' reasoning processes. Hence, by studying the team chats and relating them to the team's decision outcome, we will be able to analyze the team decision mechanism. The insights we gain in this section will also help us in Section 7 to explain our findings about team vs. individual performance.

Following the method in Cooper and Kagel (2005), we use a three-step approach to conduct the chat analysis. In the first step, we develop a coding scheme (classification of the text chats) based on the behavioral theories we want to test. The full coding scheme can be found in Appendix A.[14] Table 4 provides examples of the codes. Here we also provide an overview of the coding scheme. The coding scheme can be divided into four sub-categories:

1. **Newsvendor codes**: The codes in this sub-category are developed based on various behavioral Newsvendor decision theories. We will discuss these theories in greater detail in Section 6.1.
2. **Strategic codes**: These codes capture subjects' concrete expressions of strategic behavior, including formulating the team's own strategy, or thinking from the opponent's perspective.
3. **Trust codes**: These codes are designed to capture the trust behavior in the human retailer setting (stage 3): whether the retailer wants to be trustworthy/untrustworthy, and whether the supplier wants to be trusting/untrusting.
4. **Team dynamics**: Theses codes capture the team dynamics from round to round, including making use of the feedback and expressing regret.

Table 4 **Examples of Codes and the Identification Procedure**

Newsvendor Codes	**Identification of the Code**
Aggressive	General statements of being aggressive by producing more.
Conservative	General statements of being conservative by producing fewer.
Waste Aversion	Expressing aversion for potential waste due to over-production.
Stock-out Aversion	Expressing willingness to produce more in order to avoid stockout.
Loss Aversion	Expressing aversion for potential loss in profits.
Forecast-Dependent Risk Attitude	Being risk taking when the forecast is high, and being risk averse when the forecast is low.
Mean Anchoring	Using mean (X) as the decision benchmark and making production adjustments.
Strategic Codes	**Identification of the Code**
Own Strategy	The team discusses the general objectives of their actions/strategies.
Opponent	The team discusses how the opponent will set its strategy.
Trust Codes	**Identification of the Code**
Trustworthy Retailer	Expressing willingness to be truth-telling.
Untrustworthy Retailer	Expressing willingness to inflate.
Trusting Supplier	Expressing confidence about the retailer or the message they receive from the retailer.
Untrusting Supplier	Expressing skepticism about the message they receive from the retailer.
Team Dynamics Codes	**Identification of the Code**
Regret	The team expresses regret for what they had done in the last round/previous rounds.
Auto-correlation	Expressing Doubt about the Randomness in Demand Realization
Feedback	Referring to realized outcomes in previous rounds to facilitate decision making.

Note: The complete list of the codes is in Appendix A.

In the second step, we recruit two coders, train them with the experiment background and the coding scheme, and let them code all the chats *independently*. The two coders are quite consistent in the way they code the chats: The average cross-coder correlation for all the main codes[15] in the coding scheme equals 0.61, which is considered to be excellent consistency, according to Cooper and Kagel (2005). To better understand the coding work, consider the following dialogue from a team in the first round in the computerized retailer setting. A and B refer to the two members of the team.

[14] The complete chat data is available from the authors upon request.

[15] The "main codes" refers to all the codes that, together, account for at least 90% of the codes of that stage.

> A: “I think it’s best to pick a value near the forecast information, since the uncertainty is random.”
>
> B: “OK.”
>
> A: “Higher or lower though?”
>
> B: “Lower.”

As an example, the chat “it’s best to pick a value near the forecast information” explicitly refers to the forecast, which is the mean of demand, when making the decision; so, it is coded as *Mean Anchoring*. At the same time, in this piece of chat the team discusses and formulates a concrete Newsvendor decision strategy; hence, it will be coded as *Own Strategy*.

Finally, in the third step we examine whether the coded chat has an impact on the retailer/supplier’s decision in the current round and potentially in all the rounds. The coding we use in this step is the average of the two coder’s codings.

6.1. Chat Analysis - Computerized Retailer Setting

In this subsection, we present the study of decision mechanism in the standalone Newsvendor decision (computerized retailer setting), using chat analysis. Before we proceed, it will be useful to first present an overview of the behavioral Newsvendor decision theories. As we mentioned in the Literature Review, a very robust observation in Newsvendor experiments is that subjects often order too close to the mean, referred to as the “pull-to-center” bias. Researchers have developed different theories to explain this observation. In reviewing this literature, Thonemann and Becker-Peth (2018) conclude that there are four competing theories that give theoretical predictions consistent with the pull-to-center bias. The four theories and our related chat analysis codes are summarized below. Definition of these codes can be found in Figure 4 and Appendix A.

1. **Minimizing Ex-post Inventory Error**: This theory assumes that subjects incur psychological costs for waste and stock-out in inventory when the demand realizes. The related codes are *Waste Aversion* and *Stock-out Aversion*. They capture subjects’ explicit aversion to having waste or stock-out in inventory outcomes.

2. **Impulse Balance**: This theory assumes that waste and stock-out in inventory generate psychological impulses in opposite directions. Subjects make inventory decisions to balance the impulses. This theory is quite similar to the first one as it also highlights the psychological aspect of having waste or stock-out in inventory outcomes. Hence, the related codes are *Waste Aversion* and *Stock-out Aversion*.

3. **Quantal Choice (Bounded Rationality)**: This theory assumes while the best decision is not always made, better decisions are made more often. Most applications of the Quantal Choice theory in Operations Management typically assume heterogeneity in strategic sophistication (represented as fewer errors) across individuals and/or treatments, so looking for more concrete strategies

is a natural way to identify relatively sophisticated teams. The related code is *Own Strategy*, which captures subjects' discussion of decision strategies.

4. **Anchoring on Mean and Insufficient Adjustment Towards the Optimum**: As its name suggests, the theory assumes that people first refer to the mean of demand when making the Newsvendor decision. They know the optimal direction to adjust from the mean; but, they are unable to make the adjustment sufficiently. The related codes are *Mean Anchoring* and *Own Strategy*. *Mean Anchoring* captures subjects' explicit reference to the mean of demand when making the Newsvendor decision. *Own Strategy* captures the aspect that subjects know the optimal direction to adjust to.

Despite the fact that many different theories have been proposed, Thonemann and Becker-Peth (2018) point out that:"Various theories ... can explain the [pull-to-center] effect, but without better understanding of the cognitive processes that drive ordering behavior, it will be difficult to identify the theory or theories that best explain actual behavior." With the help of chat analysis, we are now able to shed light on this question.

There are two things that, if present in our data, can be considered empirical evidence for a theory: (1) the related codes are frequently coded; and (2) the presence of the codes impact the decision outcome as the theory suggests. In fact, we need to see *both* to conclude in favor of a theory. In Table 5, we first present the total coding frequency from teams across all 6 rounds when making the Newsvendor decision in stage 2 (computerized retailer setting). The non-integer frequency for some codes is due to the fact that we are averaging across two independent coders.

Table 5 Computerized Retailer Setting (Stage 2): Newsvendor Total Coding Frequency

Code	Frequency	Code	Frequency
Feedback	93	Own Strategy	84
Conservative	82.5	Loss Aversion	53.5
Aggressive	41.5	Auto-correlation	31.5
Mean Anchoring	30.5	Forecast-dependent Risk Attitude	20.5
Regret	16	Stock-out Aversion	0.5
Waste Aversion	0		

From Table 5, we observe that the codes *Waste Aversion* and *Stock-out Aversion* are almost never coded. Notice that *Aggressive* and *Conservative*, which are frequently coded, are not the generalization of *Waste Aversion* or *Stock-out Aversion*. *Aggressive* and *Conservative* capture subjects' mental dispositions when making Newsvendor decisions, while *Waste Aversion* and *Stock-out Aversion* capture subjects' preferences for specific operational outcomes. One demonstrative example is as follows. This piece of chat is coded as *Mean Anchoring* and *Conservative*.

A: “Should we go lower or higher?”

B: “Maybe just stay around the forecast range? Like 370.”

A: “Do we go lower to be on the safe side?”

B: “That’s good with me. Like 360?”

A: “Yeah.”

Clearly, subjects here are expressing mental dispositions in very general terms, rather than considering specific operational outcomes. To study this more formally, we use the *WordCloud* text mining package in R to analyze what people actually talk about in the chats. We find that “risk” and “optimistic” are among the most frequently mentioned words for *Agresssive* chats, and so are the words “safe” and “conservative” for *Conservative* chats; among these chats, we also find that subjects frequently refer to the mean of demand in their arguments. To put a finer point on it, if subjects were to express *Waste Aversion* or *Stock-out Aversion*, we would expect words like “waste”, “left-over”, “stock-out”, “overshoot”, or “undershoot” mentioned in the chats. However, these words are rarely mentioned in any of the chats in our experiments. In summary, the above results suggest that the operational outcomes suggested in the Minimizing Ex-post Inventory Error theory and the Impulse Balance theory are not salient in subjects’ thought processes. Meanwhile, we should also notice that the codes *Own Strategy* and *Mean Anchoring* are frequently coded. Hence, the Quantal Choice theory and the Anchoring on Mean and Insufficient Adjustment Towards the Optimum theory are still potentially consistent with our data.

Next, we consider the impact of these codes on teams’ decisions. We first focus on the code *Own Strategy* since it is a key component of both the Quantal Choice theory and the Anchoring on Mean and Insufficient Adjustment Towards the Optimum theory. Both theories assume that subjects have an explicit idea about the optimal decision. If these two theories did capture teams’ reasoning processes, teams should put effort into formulating their concrete Newsvendor decision strategies; more importantly, doing so should help them get closer to the optimum.

Interestingly, we find that talking about *Own Strategy* benefits *neither* quality *nor* consistency of teams’ Newsvendor decision-making. The correlation between each team’s total coding frequency of *Own Strategy* and the team’s average production reduction from the mean is -0.03. The correlation between each team’s total frequency of *Own Strategy* and the team’s variance of production reduction over 6 rounds is 0.10. From another angle, there are 50 teams (out of 66 teams in total) who have talked about *Own Strategy*; compared to the teams who have never talked about *Own Strategy*, the teams who have talked about *Own Strategy* have *smaller* average reduction (7.10<7.38) and *larger* reduction variance averaged across teams (1.23>1.05). Hence, it appears that teams do not benefit from concretely formulating their Newsvendor decision strategies, and they are in general unable to derive the optimal decision.[16] Based on this, the Quantal Choice theory and the Anchoring on Mean and Insufficient Adjustment Towards the Optimum theory are also inconsistent with our data.

To further validate our findings and conduct a more comprehensive analysis of the key decision drivers in Newsvendor decision-making, we consider the following regression model(s): We incorporate coding frequency as additional independent variable(s) to the regression function (1). We then consider the impact of the codes on the round that it is mentioned, i.e., the coding frequency variables are round-specific (with subscript it). Significant coefficients therefore indicate that team decisions move in a systematic direction when a particular topic is discussed. Details of the regression with chat analysis are provided in Appendix B. We consider two regression specifications: (a) include only a single chat code at a time and; (b) include all codes jointly. To give an example, when we include only the chat code *Aggressive*, we have:

$$Q_{CR,it} = \text{Intercept} + \beta_X \cdot X_t + \beta_{team} \cdot Team_i + \beta_t \cdot t + \beta_{Agg} \cdot Aggressive_{it} + \epsilon_{it}. \tag{5}$$

This specification is simply the original Newsvendor regression function (1) with the additional *Aggressive* variable, which refers to the coding frequency of the code *Aggressive* for team i in round t. Table 6 presents the results for the codes directly related to the four Newsvendor decision theories discussed above.[17]

Table 6 Computerized Retailer Setting (Stage 2): Newsvendor Chat Regression Analysis

Code	(1)	(2)	(3)	(4)	Original	Full
Own Strategy	2.60 (2.59)					4.58(2.66)*
Mean Anchoring		-0.13 (2.62)				0.12(3.20)
Aggressive			17.64 (3.35)***			15.89(3.46)***
Conservative				-6.57 (2.66)***		-7.56(2.48)***
Team Dummy Variable	6.26(3.10)**	6.82(3.01)**	4.96(2.96)*	8.17(3.08)***	6.81(2.96)**	5.57(3.25)*

Notes: Random-effects GLS regression with balanced panel data, clustering on the decision unit level. 780 observations over 6 rounds, from 66 teams and 64 individuals. The table reports estimates for the codes and the $Team$ variable. Columns 1-4 report the estimates for the specification with a single code included. Column 5 (Original) reports the estimate for the $Team$ variable from Table 1. Column 6 (Full) reports the results with all codes included. Robust standard errors are in parentheses. Significance is denoted: $^{*}p < 0.1$, $^{**}p < 0.05$,$^{***}p < 0.01$.

[16] In fact, by studying the chats we find only one team that is able to explicitly derive the optimal decision rule $Q^*_{CR} = X - 45$.

[17] Notice that we apply these regression functions to the whole data set, which includes data from both teams and individuals. The $Team$ dummy variable in these functions is used to calibrate the effect size of the coding frequency variables. For example, using the full regression in Table 6 we can conclude that the difference between a team that mentions being *Aggressive* (just once) versus one that does not is a change in production comparable to about 3 times the average individual-team difference. We also consider the alternative specification which drops the $Team$ variable, and run it only on the data from teams. The regression results are mostly unchanged.

Regression analysis provides further support that the Quantal Choice theory and the Anchoring on Mean and Insufficient Adjustment Towards the Optimum theory are inconsistent with teams' reasoning processes. Since both theories assume that subjects know the optimum, or at least its direction/approximate value, the estimates for the corresponding codes should be significantly *negative* (recall that the optimal decision is to produce *below* the mean X by 45 units). However, we see that the estimate for *Own Strategy* does not go in the direction as expected, and the estimates for *Mean Anchoring* are not significant. On the other hand, we find that the codes *Aggressive* and *Conservative* are the key decision drivers.[18] Each time a team talks about being conservative, on average, its decision goes down by 6.57 units, which brings the team closer to the optimum. Each time a team talks about being aggressive, its decision increases by 17.64 units, bringing it further away from the optimum.

So far, we find no support for any of the four theories to be the theory consistent with teams' reasoning processes. However, we are able to find support for an alternative version of the Anchoring on Mean and Insufficient Adjustment Towards the Optimum theory. Focusing on the 34 times that *Mean Anchoring* is coded, we find that teams' decisions go in both directions relative to the mean X. Moreover, we find that these decisions are closely related to teams' expressions of being conservative or aggressive.

Table 7 Computerized Retailer Setting (Stage 2): Teams' Actions when Mean Anchoring is Coded

Teams' Decisions	Frequency of *Aggressive*	Frequency of *Conservative*
Above the Mean: 10 times	4	0
Equal to the Mean: 9 times	1	3
Below the Mean: 15 times	0.5	12

Note: "Times" is the number of rounds where *Mean Anchoring* is coded at all, whereas frequency is the number of times the relevant code is mentioned in those rounds. In column 1 (2) we are reporting the *sum* of frequency that *Aggressive* (*Conservative*) is coded. For example, there are 10 times that *Mean Anchoring* is coded and subjects produce above the mean. In these 10 times, *Aggressive* is coded with a total frequency of 4 and *Conservative* is coded with a total frequency of 0.

The results are summarized in Table 7. We can see that out of the 34 times *Mean Anchoring* is coded, less than half of the time, teams choose to produce below the mean. For the 10 times teams produce above the mean, teams talk about being aggressive in those rounds and never talk about being conservative. For the 15 times teams produce below the mean, they almost never talk about

[18] As a complementary finding, the coding frequency of *Own Strategy* is positively associated with the coding frequency of *Aggressive* and *Conservative*. Hence, explicitly formulating Newsvendor decision strategies does not preclude the expression of *Aggressive/Conservative* mental dispositions.

being aggressive and often talk about being conservative. The trend of frequency dropping (from 4 to 1 to 0.5) for the code *Aggressive* and the trend of frequency increasing (from 0 to 3 to 12) for the code *Conservative* are both significant with non-parametric trend test (p-value<0.04 and p-value<0.01, respectively). Hence, we conclude that we find support for mean anchoring along with *both* aggressive and conservative adjustments. Compared to the original Mean Anchoring and Insufficient Adjustment Towards the Optimum theory, the mean anchoring part is the same, but the adjustment process is different.

An important question is: What drives the *Aggressive* and *Conservative* mental dispositions? In Appendix D, we consider several different aspects of the team's recent history, and find that three factors significantly impact the likelihood of *Aggressive* and *Conservative* arising in the following round: the coding frequency of the code in the previous round, the amount demand is above the decision in the previous round, and the realized demand uncertainty in the previous round. Additionally, we find that, in a given round teams may discuss being either *Aggressive* or *Conservative*, but they rarely discuss *both* mental dispositions in the same round, i.e., teams tend to enter one of the two mental modes (aggressive/conservative) in a particular round. These findings motivates us to connect teams' reasoning processes in Newsvendor decision-making with a behavioral economics model — the *Cue-based* decision model [Laibson (2001), Bernheim and Rangel (2004)]. The foundational assumption in the Cue-based model is that environmental cues activate different mental modes, and the subject' decision is driven by the mental mode he is in.[19] For example, for people who are addicted to nicotine, revisiting the place where he/she used to smoke or seeing an open box of cigarette will create strong cravings that trigger the smoking behavior [Laibson (2001)]. Formally, Bernheim and Rangel (2004) model this as a dynamic control problem where the subject has two mental modes: "hot" and "cold". The random environmental cue he encounters, along with past history and past actions, determines which mental mode he is in. The mental mode subsequently drives his decision in that round. In the context of Newsvendor decision-making, our findings suggest that past outcomes may serve as the cues that trigger teams to enter one of two modes: aggressive or conservative. The mode a team enters determines the team's Newsvendor decision in that round.

To summarize, in chat analysis we find no support for any of the four theories. Teams' decisions are not driven by their aversion to overstocking or under-stocking in inventory outcomes. Their effort in trying to formulate Newsvendor decision strategies do not help them get closer to the optimum, which suggests that few teams know the optimum or are able to derive it during the

[19] In some Cue-based models, the functional form of the subject's utility function is the same in different mental modes, and only the parameters are different [Laibson (2001)]. In some other Cue-based models, the subject has completely different utility functions in different mental modes [Bernheim and Rangel (2004)].

experiment. Rather, teams follow a much simpler decision heuristic in making the Newsvendor decision: they start from the mean and make both aggressive (upward) and conservative (downward) adjustment. The direction of adjustment is history dependent.

6.2. Chat Analysis - Human Retailer Setting

We now turn to the study of decision mechanism in the information sharing game (human retailer setting). Briefly recapping our results in the human retailer setting, we find that: Compared to individuals, team retailers inflate more while team suppliers reduce the same. In this section we use chat analysis to explore these observations. *Trust* and *trustworthiness* are the natural candidate drivers of these findings, as well as strategic elements including formulating one's own strategy and thinking from the opponent's perspective.

Similar to the chat analysis in the previous section for the computerized retailer setting, we first look at the coding frequency. Table 8 lists code frequencies for both the supplier and the retailer in the human retailer setting. For each, we list the top 5 most frequent codes as they cover around 90% of all the codings.

Table 8 Human Retailer Setting (Stage 3): Coding Frequency

Retailer	Frequency	Supplier	Frequency
Own Strategy	122	Own Strategy	91
Opponent	102	Opponent	82.5
Feedback	46.5	Feedback	54
Untrustworthy	65.5	Untrusting	50.5
Trustworthy	24	Trusting	22.5

Next, we consider the impact of this on the decision outcome. Similar to the chat analysis in the previous section, we include coding frequency as additional independent variables in regression functions (3) and (4).[20] The results for the retailer are summarized in Table 9.[21]

[20] On the supplier side, the coding frequency variables are round-specific, similar to the computerized retailer setting. This is motivated by the observation that *Untrusting* and *Trusting* chats spread across all six rounds. On the retailer side, however, we observe that most *Untrustworthy* and *Trustworthy* chats are in the first two rounds; that is, team retailers tend to formulate their strategies in the first few rounds they play as the retailer. Hence, for retailers, for each code we use the *summation* of coding frequency (in all the rounds the team play as the retailer) as the coding frequency variable. For completeness, in Appendix B we present both specifications (round-specific/summation) for the analysis in Table 6, 9, 10 and the R^2 and BIC comparisons; note that fit improvement is modest (and that the qualitative results are similar with the alternate specification).

[21] Similar to the computerized retailer setting, the regressions in Table 9 (and also Table 10 for the suppliers) are applied to the whole data set, which includes data from both teams and individuals; the $Team$ dummy variable here is used to calibrate the effect size of the coding frequency variables. For example, as presented in column 7 of the table, the difference between a team retailer that mentions being *Untrustworthy* (just once) versus one that does not is a change in signal decisions comparable to the average individual-team difference (11.65 vs. 13.43). We also consider the alternative specification where we only include the data from teams, and find that the regression results are mostly unchanged.

Table 9 Human Retailer Setting (Stage 3): Retailer Chat Regression Analysis

Code	(1)	(2)	(3)	(4)	(5)	Original	(7)
Untrustworthy	6.82 (3.28)**						11.65(5.02)**
Trustworthy		-5.82 (5.28)					-6.74(6.38)
Feedback			-8.86(3.90)**				-9.28(4.74)**
Own Strategy				0.44(1.94)			1.02 (2.98)
Opponent					-0.75 (1.69)		-1.71 (2.91)
Team Dummy Variable	7.86(6.63)	17.45 (6.08)***	21.99 (6.38)***	14.27(6.91)**	15.26 (6.40)**	15.13(5.71)***	13.43(7.01)*

Notes: Random-effects GLS regression with unbalanced panel data, 780 observations over 6 rounds, from 66 teams and 64 individuals. The table reports estimates for the codes and the $Team$ variable. Column 1-5 reports the estimates for the specifications with a single code included. Column 6 reports the estimate for the $Team$ variable from Table 2. Column 7 reports the results for the 5 codes and the $Team$ variable in the "full" regression. Robust standard errors are in parentheses. Significance is denoted: $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

Table 10 Human Retailer Setting (Stage 3): Supplier Chat Regression Analysis

Code	(1)	(2)	(3)	(4)	(5)	Original	Full
Untrusting	-10.97 (3.42)***						-9.54(3.73)***
Trusting		19.09 (4.89)***					20.67(5.20)***
Feedback			-6.82(3.73)*				-5.91(3.72)
Own Strategy				-5.63(3.00)*			-3.20 (2.90)
Opponent					-0.96 (2.28)		-1.00 (2.59)
Team Dummy Variable	-2.87 (4.94)	-8.07(4.85)*	-3.61 (4.99)	-2.57(5.16)	-5.17 (5.00)	-5.61(4.89)	-2.05(5.17)

Notes: Random-effects GLS regression with unbalanced panel data, 780 observations over 6 rounds, from 66 teams and 64 individuals. The table reports estimates for the codes and the $Team$ variable. Columns 1-5 report the estimates for the specifications with a single code included. The sixth column reports the estimate for the $Team$ variable from Table 3. The seventh column reports the results for the 5 codes and the $Team$ variable in the "full" regression. Robust standard errors are in parentheses. Significance is denoted: $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

We observe that although *Own Strategy* and *Opponent* are the top two most frequent codes, neither of their estimates is significant. On the other hand, the trust codes play a critical role in the decision-making process.[22] As the retailer, each time a team retailer talks about being *Untrustworthy*, on average its degree of inflation increases by 6.82 units. However, *Trustworthy* does not give a significant estimate. From the regression, we conclude that *Untrustworthy* dominates *Trustworthy* in determining retailer decisions. This helps to explain why team retailers tend to inflate more compared to individual retailers.

Why does *Trustworthy* not have a significant impact? By looking at the distribution of these two trust codes, we find that the argument of being *Untrustworthy* always wins the argument over being *Trustworthy*. Specifically, we find that many teams talk about both codes in the same round, and particularly during the early rounds of the human retailer setting. We compare the group of retailers who have talked about *both* being *Trustworthy* and *Untrustworthy* with the group of retailers who have only talked about being *Untrustworthy*. We find that the degree of

[22] As a complementary finding, none of the coding frequency of the four trust codes is correlated with the state of the game (measure by the value of X for the retailer and $\tilde{X}$ for the supplier).

inflation between these two groups are very similar (average inflation: 37 vs 35; rank-sum test on average inflation gives p-value=0.97). That is, mentioning *Trustworthy* does not have an significant impact on the retailer's strategy when *Untrustworthy* is also present, i.e., *Untrustworthy* wins the argument.

Finally, we turn to the supplier side and present the results in Table 10. We observe that similar to retailers, the main decision drivers are the trust codes rather than the strategic codes. We also observe that although *Trusting* is not as frequently coded as *Untrusting*, it is twice as impactful, and their estimates are in opposite directions. Hence, when we consider all the rounds of team suppliers, *on average,* the effect of trusting and untrusting tend to cancel out. This helps to explain why team suppliers reduce just the same as individual suppliers.

7. Discussion

In this section, we revisit the concept of the "Eureka!" type decision and discuss it in greater detail. As has been mentioned in the Introduction, there are two key characteristics of the Eureka-type decision: (1) there is a clear, unambiguous optimal solution to the question/decision task; (2) subjects within the same team are able to easily demonstrate or verify the optimality of a proposed solution, which allows them to efficiently converge to the optimal solution. Teams have been found to perform better than individuals when making the Eureka-type decision [Davis (1992), Cooper and Kagel (2005)]. The reason is that if *one team member* gets the correct answer, he/she can explain their reasoning to his/her teammates and convince them. Namely, there is a higher probability for a team to get the optimal solution compared to an individual. Both characteristics are crucial for teams to perform better than individuals. Characteristic (1) provides the premise that team members, through extensive discussions, can converge to a single solution and use it as the team's decision. Characteristic (2) ensures that the member who gets the correct solution is able to convince the whole team to follow his/her proposal, by using either logical reasoning or realized outcomes. Researchers in social psychology have found that teams make better *tactical* Eureka-type decisions compared to individuals [Lorge and Solomon (1955), Shaw (1971), Davis (1992)]. Cooper and Kagel (2005) is among the first to extend this concept to a *strategic* decision context and show that teams are more rational and strategic.

A key difference between our study and previous studies is that, unlike word puzzles or the dominant strategy in the prisoner's dilemma, in our experiments it is not straightforward to identify *ex-ante* whether the decision tasks are Eureka-type. The gap is that the decision tasks we consider may not satisfy *both* characteristic (1) and (2). Chat analysis enables us to directly study teams' reasoning process and make this identification *ex-post.* In the standalone Newsvendor decision (computerized retailer setting), the decision task has a unique optimal solution ($Q^*_{CR} = X - 45$), consistent with characteristic (1). However, finding this appropriate balance of overage and underage costs is rather complex, and its complicated nature makes it hard for subjects to demonstrate and discuss the reasoning behind their decision proposals. Chat analysis in Section 6.1 shows that teams' effort in concretely formulating Newsvendor decision strategies is uncorrelated with either the quality or consistency of their Newsvendor decisions. Namely, the teams who have spent such effort *do not* have a decision advantage over the teams who have not, and they are unable to converge to the optimum through discussions. Hence, by studying teams' reasoning processes, we conclude *ex-post* that the standalone Newsvendor decision is not a Eureka-type decision.[23]

In the human retailer setting, game theory gives a clear equilibrium prediction, consistent with characteristic (1). However, derivation of the retailer's and the supplier's optimal strategies requires different levels of cognitive sophistication. The retailer has a clear and simple strategy to "inflate from X" when reporting the signal. In contrast, the supplier needs to form high-order beliefs and use the logic of iterated dominance to derive the optimal strategy. Namely, she needs to not only form her own belief, but also consider the retailer's response to her belief (what he thinks that I think). Also, a strategy that is harder to derive in the first place will also be harder to demonstrate or verify later. This is what we have seen in the chat analysis. On the retailer side, the argument of being *Untrustworthy* dominates the argument of being *Trustworthy*, showing that team members are able to effectively communicate the benefit of "inflating more". Hence, the retailer's decision task is Eureka-type. On the supplier side, there is no uniquely compelling argument as we instead see that both the argument of *Trusting* and *Untrusting* can win conversations, which is potentially due to subjects' changing beliefs from round to round. Therefore, we conclude *ex-post* that the supplier's decision task is non-Eureka, and this is consistent with our finding that team suppliers act similarly to individual suppliers.

8. Conclusion

In this paper we study how teams make decisions in two specific inventory planning contexts: the standalone Newsvendor setting, and Newsvendor with information sharing. Past psychology

[23] This finding also helps to organize some of the earlier results on teams in the Newsvendor problem. In Gavirneni and Xia (2009), the decision proposals (which serve as potential anchors) presented to the subjects are not meant to induce subjects to make the optimal decisions, nor are they supplemented with the reasoning behind the proposals. Hence, the decision remains non-Eureka. In Laya and Pavlov (2015), one member of the team knows the optimal Newsvendor solution, but has no way to demonstrate his/her identity as the expert. Given that the Newsvendor problem is not Eureka-type, the expert member will find it difficult to demonstrate or verify the optimality of his/her decision proposal. Indeed, both papers conclude that teams perform no better than individuals. On the other hand, in Wu and Seidmann (2015), the winning individual/team's decision is shown publicly at the end of each round, and teams perform better than individuals towards the end of the experiment, suggesting that teams are better able to learn from the best *after* the superiority of the decision has been demonstrated/verified. We can understand this as follows: Teams perform better in experiments where the intervention makes the Newsvendor problem more Eureka-type.

and behavioral economics research suggest that teams should be able to make better standalone Newsvendor decisions (tactical decision-making) and should be more strategic in the Newsvendor with information sharing (strategic decision-making). Surprisingly, in this paper we find that teams are more affected by the pull-to-center bias and make *worse* Newsvendor decisions. Although the team Newsvendor decision is primarily driven by the more capable person within the team, there is loss in optimality when the two team members try to reach an agreement. Teams are also less willing to experiment with different decisions compared to individuals.[24] In the Newsvendor with information sharing, we find that team retailers are more untrustworthy than individual retailers, hence being more strategic. On the other hand, team suppliers are as trusting as individual suppliers. The notion of a "Eureka!" type decision helps us organize our findings as discussed in Section 7. Taken together, our results underscore the importance of careful behavioral studies in understanding when teams are better/worse than individuals when making business and operations decisions.

A novel feature of our study is that we use chat analysis to study the *team decision mechanism*, which has been challenging in past behavioral studies with individual subjects. Chat analysis proves to be a powerful tool to analyze whether existing decision theories are consistent with subjects' reasoning processes. In Newsvendor decision-making, we find that teams are not driven by the psychological costs of waste or stock-out in inventory; nor do they seem to hold a firm belief in the optimal direction to adjust from the mean of demand. Rather, teams employ a simple decision heuristic: they start from the mean and make both aggressive and conservative adjustments, which are history dependent. Looking forward, our results shed light on potential future research directions. The finding that teams have distinctive mental modes (aggressive/conservative) that drive their decisions is similar to the Cue-based decision model in behavioral economics. Future research can explore this mental mode pattern, but establishing a formal theoretical model or a dedicated experiment is beyond the scope of this paper.

In our experiments, we require the two members of a team to reach *mutual agreement* to form a team decision. In practice, many teams have employed such a unanimity rule because they want to respect the opinions from all the members. Of course, in practice there exists other team decision rules, for example the majority rule. This rule is not quite applicable to our experimental setting given that each team only has two members. It will be interesting to study how our results and insights can be extended to the context of large group decision-making. Another important decision

[24] Combining our results with previous studies suggests that in a typical Newsvendor setting, teams will likely perform no better than individuals, and may often perform worse. When the Newsvendor problem is made more Eureka-type (perhaps with managerial interventions), teams will be more likely to perform better. Conversely, if the decision environment is dynamic (e.g., where the mean of demand changes over time), teams will be more likely to perform worse, as teams appear to be slower at adapting and experimenting.

rule in practice is the "authority rule with discussion", where team members discuss and report their opinions to the manager, and the manager is the *only person* responsible for making the decision. This decision rule is effectively an *individual decision making* rule, while the focus of our paper is *team decision making*.

Future research can explore the implication of team decision making in other managerial and operational contexts. One possibility is to consider the operational context where the members of the same team have direct conflicts in interests and preferences, which is not the focus of this paper. Another question worth considering is de-biasing tools for teams. Incorporating de-biasing treatments and subsequently analyzing team chats can help us directly identify which existing de-biasing tool is more efficient in terms of facilitating logical reasoning; it may also help to design new de-biasing tools that target on specific behavioral biases identified in chat analysis. Being among the first paper in BOM to study teams, and given the prevalence of teams in real-world business and operations decision-making, we believe there are plenty of opportunities for future research building on our work.

References

Beer R, Ahn HS, Leider S (2017) Can trustworthiness in a supply chain be signaled? forthcoming. *Management Science* .

Bernheim BD, Rangel A (2004) Addiction and cue-triggered decision processes. *American economic review* 94(5):1558–1590.

Bolton GE, Katok E (2008) Learning by doing in the newsvendor problem: A laboratory investigation of the role of experience and feedback. *Manufacturing & Service Operations Management* 10(3):519–538.

Bolton GE, Ockenfels A, Thonemann UW (2012) Managers and students as newsvendors. *Management Science* 58(12):2225–2233.

Bostian AA, Holt CA, Smith AM (2008) Newsvendor pull-to-center effect: Adaptive learning in a laboratory experiment. *Manufacturing & Service Operations Management* 10(4):590–608.

Charness G, Frechette GR, Kagel JH (2004) How robust is laboratory gift exchange? *Experimental Economics* 7(2):189–205.

Charness G, Karni E, Levin D (2007) Individual and group decision making under risk: An experimental study of bayesian updating and violations of first-order stochastic dominance. *Journal of Risk and uncertainty* 35(2):129–148.

Charness G, Karni E, Levin D (2010) On the conjunction fallacy in probability judgment: New experimental evidence regarding linda. *Games and Economic Behavior* 68(2):551–556.

Charness G, Sutter M (2012) Groups make better self-interested decisions. *The Journal of Economic Perspectives* 26(3):157–176.

Cooper DJ, Kagel JH (2005) Are two heads better than one? team versus individual play in signaling games. *The American economic review* 95(3):477–509.

Cox JC (2002) Trust, reciprocity, and other-regarding preferences: Groups vs. individuals and males vs. females. *Experimental Business Research*, 331–350 (Springer).

Davis JH (1992) Some compelling intuitions about group consensus decisions, theoretical and empirical research, and interpersonal aggregation phenomena: Selected examples 1950–1990. *Organizational Behavior and Human Decision Processes* 52(1):3–38.

Donohue K, Katok E, Leider S (2018) *The Handbook of Behavioral Operations* (Wiley).

Fehr E, Kirchsteiger G, Riedl A (1998) Gift exchange and reciprocity in competitive experimental markets. *European Economic Review* 42(1):1–34.

Fischbacher U (2007) z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics* 10(2):171–178.

Gavirneni S, Xia Y (2009) Anchor selection and group dynamics in newsvendor decisionsa note. *Decision Analysis* 6(2):87–97.

Greiner B (2004) An online recruitment system for economic experiments. Technical report, University Library of Munich, Germany.

Ho TH, Lim N, Cui TH (2010) Reference dependence in multilocation newsvendor models: A structural analysis. *Management Science* 56(11):1891–1910.

Katok E, Wu DY (2009) Contracting in supply chains: A laboratory investigation. *Management Science* 55(12):1953–1968.

Kocher MG, Sutter M (2005) The decision maker matters: Individual versus group behaviour in experimental beauty-contest games. *The Economic Journal* 115(500):200–223.

Kremer M, Moritz B, Siemsen E (2011) Demand forecasting behavior: System neglect and change detection. *Management Science* 57(10):1827–1843.

Kremer M, Siemsen E, Thomas DJ (2015) The sum and its parts: Judgmental hierarchical forecasting. *Management Science* 62(9):2745–2764.

Kugler T, Bornstein G, Kocher MG, Sutter M (2007) Trust between individuals and groups: Groups are less trusting than individuals but just as trustworthy. *Journal of Economic psychology* 28(6):646–657.

Laibson D (2001) A cue-theory of consumption. *The Quarterly Journal of Economics* 116(1):81–119.

Laya N, Pavlov V (2015) Team decision-making and individual learning in the newsvendor problem .

Lorge I, Solomon H (1955) Two models of group behavior in the solution of eureka-type problems. *Psychometrika* 20(2):139–148.

Ockenfels A, Selten R (2014) Impulse balance in the newsvendor game. *Games and Economic Behavior* 86:237–247.

Özer Ö, Zheng Y, Chen KY (2011) Trust in forecast information sharing. *Management Science* 57(6):1111–1137.

Özer Ö, Zheng Y, Ren Y (2014) Trust, trustworthiness, and information sharing in supply chains bridging china and the united states. *Management Science* 60(10):2435–2460.

Salvi C, Bricolo E, Kounios J, Bowden E, Beeman M (2016) Insight solutions are correct more often than analytic solutions. *Thinking & reasoning* 22(4):443–460.

Schweitzer ME, Cachon GP (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Science* 46(3):404–420.

Shaw ME (1971) Group dynamics: The psychology of small group behavior .

Su X (2008) Bounded rationality in newsvendor models. *Manufacturing & Service Operations Management* 10(4):566–589.

Thonemann U, Becker-Peth M (2018) Behavioral inventory decisions: The newsvendor and other inventory settings. *The Handbook of Behavioral Operations* .

Tversky A, Kahneman D (1983) Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review* 90(4):293.

Wan Z, Beil DR, Katok E (2012) When does it pay to delay supplier qualification? theory and experiments. *Management Science* 58(11):2057–2075.

Wu T, Seidmann A (2015) Are groups better than individuals at making decisions? an experimental study on newsvendor problem. *System Sciences (HICSS), 2015 48th Hawaii International Conference on*, 4252–4261 (IEEE).

Appendix A: The Coding Scheme

In this appendix we provide the complete coding scheme document we gave to the coders. We trained the two coders separately. For each coder, we set up two 2-hour training sessions and we allowed him/her to ask any questions during the coding work. The two coders never had the opportunity to interact with each other. We use the average of their coding results (the coding frequency) as the data input to our chat analysis.

Supply Chain Coding Scheme

We have divided the coding scheme for the supply chain game into the following three sub-categories:

1. Strategic Behavior

OS: Talking about Own Strategy

The team discusses the general objectives of their actions/strategies. The directional statements should be included (go higher/lower). However, non-tactical number proposing statements should not be coded.

Example: “As the supplier, I think we can do -50 this round to be conservative.”

OPP: Thinking from the Opponent’s Perspective

The team discusses how the opponent will set the strategy, or how the opponent will respond to the team’s action/strategy.

Example 1: “I think the retailer will always go high by a large amount.”

Example 2: “ They will produce more when we pick a higher value, so let’s go high. ”

2. Trust Behavior

TRQ: Raising Questions Regarding Being Trustworthy or Not as the Retailer

The team raises an **open question** regarding whether they should be trustworthy or not as the retailer.

Example: “Do you think we should tell them the truth? ”

TR: Expressing Willingness to be Trustworthy as the Retailer

Expressing willingness to be truth-telling.

Example: “At least we are not fooling them.”

UTR: Expressing Willingness to be Untrustworthy as the Retailer

Expressing willingness to inflate, potentially by a large amount.

Example: “We want them to produce as many as possible, so its best for us to give a higher prediction.”

TSQ: Raising Questions Regarding Being Trusting or Not as the Supplier

As the supplier, the team raises an **open question** regarding whether they should trust the retailer or not.

Example: “Do you think we should trust the retailer?”

TS: Expressing Willingness to be Trusting as the Supplier

Expressing confidence for their opponent or the message they receive.

Example: “I think we can trust the retailer.”

UTS: Expressing Willingness to be Untrusting as the Supplier

Expressing skepticism for the message they receive.

Example: “I’m not trusting them, definitely lying.”

3. Team Dynamics

REG: Expressing Regret

The team expresses regret for what they had done in the last round/past few rounds. In particular, the team discusses **counterfactual** situations.

Example: “We should have gone even higher as the retailer last round.”

FB: Using Feedback from Last Round

Referring to realized outcomes in previous rounds **to facilitate decision making for the current round**.

Example: “So again, same strategy, go high when we report since that worked last time”

FORW: Forward Looking and Formulating Strategies in Advance for Being in the Same Role

Discussing what the team should do if they play **in the same role** in future rounds.

Example: “Let’s go even lower if we are **still** the supplier next round.”

FLIP: Formulating Strategies in Advance for Being in the Different Role

The team discusses strategies for being in the **different** role in future rounds.

Example: “We inflate a lot as the retailer this round. So in the future if we are the supplier, we should be super conservative.”

RAND (Auto-correlation): Expressing Doubt about the Randomness in Demand Realization

The team expresses doubt about the random nature of demand realization.

Example: "Demand in the last few rounds suggests that there's a pattern."

Newsvendor Coding Scheme

RS: Risk Seeking/Aggressive[25]

General statements of taking risk by producing more.

Example: "High risk high reward."

RA: Risk Aversion/Conservative

General statements of trying to be safe by producing fewer.

Example: " If we overshoot we will lose a lot of money, so let's play it safe."

WA: Waste Aversion

Expressing aversion for potential waste due to over-production.

Example: "I don't want to have leftovers in production, so let's go low."

SOA: Stock-out Aversion

Expressing willingness to produce more in order to avoid stockout.

Example:"I think we should go high. If we don't produce enough we may not be able to fulfill all the demand."

LA: Loss Aversion

Expressing aversion for potential **loss in profits**.

Example: "If we do so it's guaranteed that our profit won't go negative."

ANCM: Anchoring on Mean(X) and Insufficient Adjustment

Using mean (X) as the decision benchmark and making production adjustments.

Example: "I think we can just pick a number around the forecast information in each round, say from -10 to +10."

FDRA: Forecast-Dependent Risk Attitude

Being risk taking when the forecast is high, and being risk averse when the forecast is low.

Example: "Even if we overproduce, it won't matter as much if the forecast is high."

General Coding Rules

1) Notice that the codes apply for all the chats. For example, in the Newsvendor task (computerized retailer setting), subjects may discuss own strategy. In the supply chain game, the supplier may discuss how to make Newsvendor-type production decisions.

2) In principle we can code one chat with multiple codes. However, be careful when attaching codes. If you are unsure about how to code some chat, code it in the way that you think most other people will agree with.

3) In the chats for both the supply chain game and the Newsvendor task (computerized retailer setting), you will see lots of number proposing. Do not code the chats that only proposed numbers. Code them when the subject expresses the reason for doing so.

[25] In the main text RS and the following RA are simply referred to as *Aggressive* and *Conservative*.

For example, if the chat is "How about 150?", do not code this chat. If the chat is "Let's do 150 to be conservative this round," code it accordingly.

4) Sometimes, subjects engage in conversations about an action/strategy. In this case, do not assign codes if the subject is only repeating his/her teammate's proposal. Code it when this chat adds something new to the discussion.

Also, sometimes a subject uses several chats to express his/her own opinion. In this case, code the chat only when the subject is adding something new in that chat rather than repeating his own opinion.

Appendix B: Details of the Regression for Chat Analysis

1. Regression Function Specifications

As we mentioned in Section 6, we consider two regression specifications for chat analysis: the regression with a single code as the additional independent variable, and the regression with all the codes included. Here, we give more details for these two specifications. For the Newsvendor chat analysis, the original regression function (1) is:

$$Q_{CR,it} = \text{Intercept} + \beta_X \cdot X_t + \beta_{team} \cdot Team_i + \beta_t \cdot t + \epsilon_{it}.$$

In the specification with a single code included, we consider the coding frequency of the codes *one at a time* as the additional independent variable. For example, the regression with the code *Conservative* is given as:

$$Q_{CR,it} = \text{Intercept} + \beta_X \cdot X_t + \beta_{team} \cdot Team_i + \beta_t \cdot t + \beta_{Cons} \cdot Conservative_{it} + \epsilon_{it}.$$

In the specification with all the codes included, we consider all the codes with frequency greater than 10. The regression function is given as:

$$\begin{aligned} Q_{CR,it} =& \text{Intercept} + \beta_X \cdot X_t + \beta_{team} \cdot Team_i + \beta_t \cdot t + \beta_{Cons} \cdot Conservative_{it} \\ &+ \beta_{Agg} \cdot Aggressive_{it} + \beta_{FB} \cdot Feedback_{it} + \cdots + \epsilon_{it}. \end{aligned}$$

The same method applies to the chat analysis in the human-retailer setting: the coding frequency of the codes are added as the additional independent variables to the regression function (3) and (4). For example, the specification with all the codes included on the supplier side is given as:

$$\begin{aligned} Q_{HR,it} =& \text{Intercept} + \beta_{signal} \cdot \tilde{X}_{it} + \beta_{team} \cdot Team_i + \beta_{NV} \cdot (\overline{X - Q_{CR}})_i + \beta_t \cdot t + \beta_{OS} \cdot OwnStrategy_{it} \\ &+ \beta_{OPP} \cdot Opponent_{it} + \beta_{UTS} \cdot Untrusting_{it} + \beta_{TS} \cdot Trusting_{it} + \beta_{FB} \cdot Feedback_{it} + \epsilon_{it}. \end{aligned} \tag{6}$$

2. Coding Frequency as Additional Independent Variables: Round-Specific or Summation?

Notice that for the supplier chat analysis, in both the computer retailer setting and the human retailer setting, the subscript for the coding frequency of the codes is *it*. In other words, we consider the impact of the code only in that round when it is coded. This is driven by the observation that in these cases, the coded chats spread across all 6 rounds. This suggests that as suppliers teams are very responsive to new information and feedback they receive in each round, instead of formulating a strategy at the beginning of the task and sticking to it.

However, the situation is different on the retailer side in the human retailer setting. We observe that most of the coded chats are in the first two rounds. This suggests that the retailers tend to formulate strategies at the beginning of the game. Moreover, recall that the retailers inflate more over time (the estimate for the time trend variable t is significantly positive). In subsequent rounds, the retailers stick to the strategies they have formulated in the early rounds, and they do so without having to refer to these strategies again. Driven by this observation, in the retailer chat analysis, the subscript for coding frequency of the codes is i. That is, we consider the impact of the codes on *all the rounds* that the team acts as the retailer, rather than for the specific round that the code is mentioned.

To give an example, the specification with all the codes included on the retailer side is given as:

$$\begin{aligned}\tilde{X}_{it} =& \text{Intercept} + \beta_X \cdot X_t + \beta_{team} \cdot Team_i + \beta_t \cdot t + \beta_{OS} \cdot OwnStrategy_i + + \beta_{OPP} \cdot Opponent_i \\ &+ \beta_{UTR} \cdot Untrustworthy_i + \beta_{TR} \cdot Trustworthy_i + \beta_{FB} \cdot Feedback_i + \epsilon_{it}.\end{aligned} \tag{7}$$

Here, for example, the variable $Untrustworthy_i$ refers to the *summation* of the coding frequency of the code *Untrustworthy* for team i as the retailer.

For completeness, here we present here the chat analysis under both specifications (round-specific or summation). Both R^2 and BIC are in favor of the specification we choose in the main text.[26] It is also important to note that while our preferred specifications have in general more precise estimates, the overall picture is qualitatively similar between both specifications.

Table 11 Computerized Retailer Setting Supplier Chat Regression Analysis: Round Specific vs Summation

Code	Round-Specific	Summation
Feedback	1.12(2.49)	3.14(1.05)***
Own Strategy	4.58(2.66)*	1.38(1.62)
Loss Aversion	-4.04(2.78)	-2.34(2.08)
Mean Anchoring	0.12(3.20)	-3.61(2.33)
Aggressive	15.89(3.46)***	2.27(1.64)
Conservative	-7.56(2.48)***	-4.55(1.36)***
Auto-correlation	7.54(2.47)***	3.14(2.25)
Forecast-Dependent Risk Attitude	0.20(4.50)	3.22(4.02)
Regret	-3.46(5.53)	4.43(4.15)
Team	5.57(3.25)*	4.88(3.83)
R^2	0.96	0.96
BIC	7216.01	7232.84

Notes: Random-effects GLS regression by including the frequency of all the codes in the table as additional independent variables to function (1). Clustering is at the decision unit level. Robust standard errors in parentheses. Significance is denoted: $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

[26] The BIC in Table 11-13 are derived by running the maximum likelihood estimation of the corresponding regression models.

Table 11 shows that using round-specific chat variables improves the precision of the estimates and increases the number of chat codes that are informative about supplier behavior with a computerized retailer. This is consistent with our observation that suppliers tend to decide based on the new information and feedback they receive in *each round*.

Table 12 Human Retailer Setting Retailer Chat Regression Analysis: Round Specific vs Summation

Code	Round-Specific	Summation
Untrustworthy	3.15(4.30)	11.65(5.02)**
Trustworthy	-2.31(5.37)	-6.74(6.38)
Feedback	-0.46(4.36)	-9.28(4.74)**
Own Strategy	0.36(2.97)	1.02 (2.98)
Opponent	1.37(2.65)	-1.71 (2.91)
Team	13.48(5.92)**	13.43(7.01)*
R^2	0.67	0.69
BIC	3838.72	3824.70

Notes: Random-effects GLS regression by including the frequency of all the codes in the table as additional independent variables to function (3). Robust standard errors in parentheses. Significance is denoted: $^{*}p < 0.1$, $^{**}p < 0.05$,$^{***}p < 0.01$.

Table 12 shows that using the summation of frequency as chat variables gives more precise estimates and increases the number of chat codes that are informative about retailer behavior. This is consistent with our observation that retailers tend to formulate the strategies they will use throughout the game in early rounds (typically in the first round they play as the retailer).

Table 13 shows that using round-specific chat variables improves the precision of the estimates and increases the number of chat codes that are informative about supplier behavior with a human retailer. This is consistent with our observation that suppliers tend to condition their expression of *Trusting* or *Untrusting* on the new information and feedback they receive in each round.

Appendix C: The Opponent Effect in the Human Retailer Setting

In the main text, we focus on the behavior of the *decision maker* while not paying much attention to the identity (team or individual) of the opponent. In this appendix, we show that our results in the main text are robust to the identity of the opponent.

1. Regression Analysis Conditioning on the Opponent - Decision Comparison

Table 14 presents the opponent effect on the retailers' degree of inflation from the demand forecast they receive. Table 15 summarizes the suppliers' degree of reduction from the signals they receive. In the third row in both tables, we run the regression functions in the main text (function (3) for the retailers and function (4) for the suppliers), *conditioning on the identity of the opponent.*

Table 13 Human Retailer Setting Supplier Chat Regression Analysis: Round Specific vs Summation

Code	Round-Specific	Summation
Untrusting	-9.54(3.73)***	-2.81(3.28)
Trusting	20.67(5.20)***	10.47(5.76)*
Feedback	-5.91(3.72)	-0.49(4.00)
Own Strategy	-3.20 (2.90)	-2.15(3.03)
Opponent	-1.00 (2.59)	-1.42(3.25)
Team	-2.05(5.17)	-1.50(6.04)
R^2	0.79	0.79
BIC	3271.14	3745.62

Notes: Random-effects GLS regression by including the frequency of all the codes in the table as additional independent variables to function (4). Robust standard errors in parentheses. Significance is denoted: ${}^{*}p < 0.1$, ${}^{**}p < 0.05$,${}^{***}p < 0.01$.

Table 14 Human Retailer Setting (Stage 3): Retailer Signal Inflation

	Average of Signal Inflation	
The Decision Maker	Playing Against Individual Suppliers	Playing Against Team Suppliers
Team Retailers	35.95	22.14
Individual Retailers	6.37	20.00
Estimate for β_{Team}	26.25(8.58)***	3.64(7.70)

Notes: Significance is denoted: ${}^{*}p < 0.1$,${}^{**}p < 0.05$,${}^{***}p < 0.01$.

Table 15 Human Retailer Setting (Stage 3): Supplier Production Reduction from Signal

	Average of Production Reduction	
The Decision Maker	Playing Against Individual Retailers	Playing Against Team Retailers
Team Suppliers	44.39	29.38
Individual Suppliers	22.5	39.60
Estimate for β_{Team}	-24.64(6.94)***	11.42(6.49)*

Notes: Significance is denoted: ${}^{*}p < 0.1$,${}^{**}p < 0.05$,${}^{***}p < 0.01$. Notice that a negative estimate means a *larger* reduction.

When fixing the identify of the opponent to be *individual*: (1) team retailers inflate significantly more compared to individual retailers; (2) team suppliers reduce significantly more compared to individual suppliers. Namely, when playing against individuals, teams are both more untrustworthy and more untrusting, consistent with what we have conjectured in Hypothesis 3. With this, it is natural to ask whether the team effect will "cancel out" when we have teams on both sides. When fixing the identify of the opponent to be *team*, we find that: (1) team retailers inflate similarly to individual retailers; (2) team suppliers reduce significantly less compared to individual suppliers. Namely, when playing against teams, team retailers are no more untrustworthy and team suppliers are even more trusting.

We will not be able to draw a complete conclusion without making necessary connections between Table 14 and Table 15. For example, individual suppliers reduce a very large amount when playing against team retailers (39.60 units); however, this may be driven by the large inflation amount from team retailers when playing against individual suppliers (35.95 units), and the *net effect* is unclear. To be able to understand whether teams have a decision advantage over individuals, we need to compare the *profits* they earn in different scenarios.

2. Regression Analysis Conditioning on the Opponent - Profit Comparison

To do this, we consider the following simple regression function:

$$Profit_{it} = \text{Intercept} + \beta_{team} \cdot Team_i + \beta_X \cdot X_t + \epsilon_{it}.$$

We can apply this function to both the retailers and the suppliers. Notice that we need to control for the demand forecast X_t because while X_t is the same for all subjects in a particular round, it varies from round to round; hence, we include it to control for the variation in profits caused by the changing market conditions.

Table 16 **Human Retailer Setting (Stage 3): Retailer Profit Comparison Regression Analysis**

	Estimate of the Coefficient		
Variable	Playing Against Individual Suppliers	Playing Against Team Suppliers	Pooling
Team	315.15(277.80)	589.00(253.21)**	459.71(169.24)***
X_t	31.54(1.70)***	30.66(1.76)***	31.04(1.22)***
Intercept	1383.21(520.42)***	1386.40(553.81)**	1386.62(378.72)***

Notes: Random-effects GLS regression with unbalanced panel data. Robust standard errors are in parentheses. 390 observations from 122 decision units for the pooling data. Significance is denoted: $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

Table 17 **Human Retailer Setting (Stage 3): Supplier Profit Comparison Regression Analysis**

	Estimate of the Coefficient		
Variable	Playing Against Individual Retailers	Playing Against Team Retailers	Pooling
Team	-64.32(229.91)	84.82(313.74)	4.92(180.60)
X_t	10.27(2.03)***	13.91(2.41)***	11.90(1.58)***
Intercept	1396.62(612.01)**	297.05(702.26)	899.69(466.00)*

Notes: Random-effects GLS regression with unbalanced panel data. Robust standard errors are in parentheses. 390 observations from 121 decision units for the pooling data. Significance is denoted: $^{*}p < 0.1, ^{**}p < 0.05, ^{***}p < 0.01$.

The results are presented in Table 16 and Table 17. From Table 16, we observe that overall team retailers earn significantly higher profits compared to individual retailers. Hence, teams do have a decision advantage over individuals when acting as the retailer. Moreover, such an advantage comes mainly from the situation when playing against team suppliers. From Table 17, we observe that in all three scenarios team suppliers earn similar profits compared to individual suppliers. Hence, teams do not have a decision advantage over individuals when acting as the supplier.

Appendix D: Analysis of the drivers of *Aggressive* and *Conservative* mental dispositions in the Computerized Retailer Setting

For each code, we consider a random-effects model that regresses its coding frequency on three variables: its coding frequency in the last round, the ξ value in the last round (realized demand uncertainty), and the

Table 18 Human Retailer Setting (Stage 3): Retailer Profit Comparison Regression Analysis

	Estimate of the Coefficient		
Code	Corresponding Code's Frequency in the Last round	$(D_{t-1} - Q_{CR,t-1})$	ξ_{t-1}
Aggressive	0.2832(0.1311)**	-0.0010(0.0005)**	0.0008(0.0005)
Conservative	0.1105(0.0542)**	-0.0009(0.0008)	0.0023(0.0010)**

Notes: Random-effects GLS regression with unbalanced panel data. Robust standard errors are in parentheses. 330 observations from 66 teams. Significance is denoted: $^{*}p < 0.1,^{**}p < 0.05,^{***}p < 0.01$.

amount demand D is above the production decision Q_{CR} in the last round. The results are summarized in Table 18.

For each of *Aggressive* and *Conservative*, its coding frequency is positively assotiated with its own coding frequency in the last round. Higher $(D_{t-1} - Q_{CR,t-1})$ discourages *Aggressive* in the following round, while higher ξ_{t-1} encourages *Conservative* in the following round. In general, the results suggest that aggressiveness and conservatism are influenced by the recent history of the decision task.

We also consider the *distribution* of the codes *Aggressive* and *Conservative* to answer two questions. (a) Are such mental dispositions driven by individual heterogeneity, i.e., do we see aggressive teams and conservative teams? (b) Do teams talk about being *Aggressive* and *Conservative* in the same round or in different rounds? To answer (a), we calculate the correlation between the coding frequency in the first three rounds and the coding frequency in the last three rounds, and find it to be 0.27 for *Aggressive* and 0.30 for *Conservative*. Namely, it is *not* the case that some teams keep expressing *Aggressive/Conservative*.[27] To answer (b), we count the frequency where *Aggressive* and *Conservative* are coded in the same round. Out of the 87 rounds *Conservative* is coded, 69 of them do not have *Aggressive* coded in the same round; out of the 52 rounds *Aggressive* is coded, 34 of them do not have *Conservative* coded in the same round. Hence, within one round teams tend to enter *one of* two mental modes: aggressive or conservative. In addition, from *WordCloud* text mining analysis, we find that the word "time" is among the most frequently coded words in *Aggressive* and *Conservative* chats; the chat that has time is typically "we should do ... this time." This demonstrates that, for mental dispositions, history is one activating factor, and when activated the aggressive/conservative disposition is framed as what to do that round (not what to do for the whole rest of the task).

[27] We also consider whether teams expressing both Aggression and Conservatism merely reflect internal disagreement between an aggressive team member and a conservative one. However, out of the teams that were coded as both *Aggressive* and *Conservative* (at least once for each in 6 rounds), 60% of teams had at least one of the codes coming from both team members.