Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Bounded Rationality in Service Systems

Tingliang Huang, Gad Allon, Achal Bassamboo,

To cite this article:
Tingliang Huang, Gad Allon, Achal Bassamboo, (2013) Bounded Rationality in Service Systems. Manufacturing & Service Operations Management 15(2):263-279. http://dx.doi.org/10.1287/msom.1120.0417

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 15, No. 2, Spring 2013, pp. 263–279
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.1120.0417

Bounded Rationality in Service Systems

Tingliang Huang
Department of Management Science and Innovation, University College London,
London WC1E 6BT, United Kingdom, t.huang@ucl.ac.uk

Gad Allon, Achal Bassamboo
Kellogg School of Management, Northwestern University, Evanston, Illinois 60208
{g-allon@kellogg.northwestern.edu, a-bassamboo@kellogg.northwestern.edu}

The traditional operations management and queueing literature typically assumes that customers are fully rational. In contrast, in this paper we study canonical service models with boundedly rational customers. We capture bounded rationality using a model in which customers are incapable of accurately estimating their expected waiting time. We investigate the impact of bounded rationality from both a profit-maximizing firm's perspective and a social planner's perspective. For visible queues with the optimal price, bounded rationality results in revenue and welfare loss; with a fixed price, bounded rationality can lead to strict social welfare improvement. For invisible queues, bounded rationality benefits the firm when its level is sufficiently high. Ignoring bounded rationality, when present yet small, can result in significant revenue and welfare loss.

Key words: behavioral operations; service operations; bounded rationality; queueing; consumer behavior
History: Received: June 21, 2011; accepted: September 1, 2012. Published online in *Articles in Advance* March 1, 2013.

1. Introduction and Literature Review

When a customer calls a call center or goes to a fast-food restaurant, a café, or an ATM and has to queue for service, does he always *accurately* and *perfectly* calculate the benefits and costs of joining before making his decisions? The traditional economics and queueing literature has assumed that he does, whereas anecdotal evidence and experimental studies point to the contrary. In this paper we study queueing or service systems without making this "perfect rationality" assumption on the part of customers. Our research questions are the following: How should one model customer bounded rationality in service systems? What are the implications of bounded rationality, for example, on firm revenue, social welfare, and pricing?

Naor (1969) appears to be the first to incorporate customer decisions into a queueing model. Naor and subsequent researchers following his work assume customers to be fully rational and able to *perfectly* estimate their expected waiting time and thus the expected utility of joining. Naor (1969) shows that self-interested customers would join a more congested system than what the social planner prescribes, and he proposes "levying tolls" (i.e., pricing) as a way to maximize social welfare. In Naor's model, customers are assumed to be able to compute with great precision the expected waiting time and thus the expected utility they are about to obtain from making a decision about whether to join or renege. One may ask, are customers fully rational? Specifically, does a customer necessarily have the capability to perfectly estimate his expected waiting time and utility? Ariely (2009) claims that irrationality is the real invisible hand that drives human decision making. Indeed, there is abundant empirical evidence that people are boundedly rational. In this work, we study the effects and implications of bounded rationality in canonical queueing or service systems.

Our study is related to several branches of the literature: economics of queues, bounded rationality in economics, and behavioral operations.

Economics of Queues. Naor (1969) studies the economics of queueing systems when customers are fully rational. Yechiali (1971, 1972) extends Naor's model to allow for $GI/M/1$ queues. Knudsen (1972) extends Naor's model to allow for a multiserver queueing system in which arriving customers' net benefits are heterogeneous. Lippman and Stidham (1977) extend the Naor model to the finite-horizon and discounted cases, showing that, in these settings, the economic notion of an external effect has a precise quantitative interpretation. Hassin (1986) considers a revenue maximizing server who has the opportunity to suppress information on actual queue length, leaving customers to decide whether to join the queue on the basis of the known distribution of waiting times. See Van Mieghem (2000), Hassin and Haviv (2003), Afèche (2004), and Hsu et al. (2009) for other extensions and a comprehensive literature review. Although various models along this line are studied, one common theme in this literature is that full rationality is always assumed.

Bounded Rationality in Economics. Traditional economic theory postulates that decision makers are "rational"; i.e., they have sufficient abilities to do perfect optimization in their choices. Simon (1955) seems to be the first to propose an alternative way to model decision-making behavior: rather than optimizing perfectly, agents search over the alternatives until they find "satisfactory" solutions. Simon (1957) coins the term "bounded rationality" to describe such human behavior.

Bounded rationality refers to a variety of behavioral phenomena in the literature. For a description of systematic errors made by experimental subjects, see Arkes and Hammond (1985), Hogarth (1980), Kahneman et al. (1981), Nisbett and Ross (1980), and the survey papers by Payne et al. (1992) and Pitz and Sachs (1984). Tversky and Kahneman (1974) show that people rely on a limited number of heuristic principles that in general are useful but sometimes lead to severe and systematic errors. On the basis of the evidence, Conlisk (1996) offers four convincing reasons for incorporating bounded rationality in economic models. Geigerenzer and Selten (2001) adopt heuristics or rules of thumb to model bounded rationality. Thurstone (1927) and Luce (1959) appear to be the first to develop the framework for *stochastic* choice rules capturing that better options are chosen more often. This approach has attracted considerable attention and has been adopted in a variety of settings. For example, following this approach, McKelvey and Palfrey (1995) and Chen et al. (1997) develop a new equilibrium concept quantal response equilibrium in game theory. Others include, for example, Bajari and Hortacsu (2001, 2005) in auctions, Cason and Reynolds (2005) in bargaining, Basov (2009) in monopolistic screening, and Waksberg et al. (2009) in natural environments. We also adopt this approach in the paper, in the context of expected waiting time estimation of service systems.

Behavioral Operations. Gino and Pisano (2008) survey the literature on modeling bounded rationality in economics, finance, and marketing, and they argue that operations management scholars should incorporate departures from the rationality assumption into their models and theories. There is an emerging literature on behavioral operations management: Lim and Ho (2007) and Ho and Zhang (2008) conduct experiments on designing pricing contracts for boundedly rational customers, Davis (2011) investigates pull contracts in controlled experiments, and Kremer et al. (2011) analyze how individuals make forecasts based on time-series data and find that forecasting behavior systematically deviates from normative predictions. We refer readers to Bendoly et al. (2006, 2009) for this stream of research. We point out two papers that are closely related to our work. The first paper is by Su (2008), who studies bounded rationality in operational settings. He applies the logit choice framework to the classic newsvendor model and characterizes the ordering decisions made by a boundedly rational (i.e., noisy) decision maker. He identifies systematic biases and investigates the impact of these biases on several operational settings. We apply a similar framework in this paper, but we interpret bounded rationality as the incapability of estimating expected waiting time in a service setting. There are several notable differences between Su (2008) and our study: First, Su's (2008) model is static (i.e., one-shot), whereas our model is a dynamic one. Second, the externality among boundedly rational decision makers is prominent in our setting, whereas it is not in Su (2008). Finally, different from Su (2008), where there are prior empirical and experimental studies of the newsvendor model that allow for statistical tests, our study is theoretical and aims to obtain testable theoretical predictions that stimulate future empirical and experimental work in service systems. Recently, Kremer and Debo (2012) have presented experimental findings along this line. The second paper is by Plambeck and Wang (2010), who study implications of hyperbolic discounting in service systems. Although the research setting (i.e., service systems) is similar, the research focus and approach are quite different. In their paper, customers lack the self-control to undergo an unpleasant experience that would be in their long-run self-interest, which is modeled by psychologists in terms of a hyperbolic discount rate for utility. Our model of bounded rationality focuses on customer ability to compute the expected waiting time.

Another stream of research related to ours is experimental study in queueing. This literature does not support that individuals are fully rational. Rapoport et al. (2004) study a class of queueing problems with endogenous arrival times formulated as noncooperative n-person games in normal form. Results from their experimental study cannot be fully explained by rational behavior; see also Bearden et al. (2005) and Seale et al. (2005) for studies along this line.

Traditional models in operations assumed that customers are rational both in maximizing their utility and in their ability to compute the anticipated expected waiting time in high precision, regardless of the complexity of the system. In this paper, our model of bounded rationality focuses on this ability to predict expected waiting times. In particular, we assume that customers may lack the capability to accurately estimate their expected waiting time (including the service time) before making their join or balk decisions and that this capability may vary across different queue configurations (e.g., visible and invisible). We thus introduce a random error term into customer's expected waiting time estimation that

reflects their inability to accurately assess the utility obtained from each action. From both a revenue-maximizing firm's perspective and a social planner's perspective, we study the impact of bounded rationality for both visible queues (such as a fast-food restaurant, a café, or an ATM) and invisible queues (such as a call center).

The three main contributions of our study are the following: (i) ours is among the first to model bounded rationality in service systems, for both visible and invisible queues; (ii) our study provides insights on how service systems should be managed in the presence of boundedly rational customers (e.g., we show that when the system manager can reduce the ambiguity associated with the process of estimating the expected waiting time, completely eliminating consumer bounded rationality may not be the optimal strategy and ignoring bounded rationality may result in significant revenue and social welfare loss); and (iii) we provide a framework to stimulate future empirical and experimental work in service systems and behavioral operations.

The remainder of this paper is organized as follows. Section 2 presents a model of bounded rationality in service systems. We study revenue maximization and social welfare maximization in §§3 and 4. We provide a discussion in §5. Proofs of the results are relegated to the online supplement (available at http://dx.doi.org/10.1287/msom.1120.0417).

2. A Model of Bounded Rationality in Service Systems

Consider a customer who has to decide whether to join a service system or not. If he joins, he will obtain expected utility $U_1 \equiv R - p - C\,\mathbb{E}\,w$, where $R > 0$ is the reward on completion of service, p is the price, $\mathbb{E}\,w$ is the expected waiting and service time, and $C > 0$ is the average waiting and service cost per unit of time. If he balks, he will get utility $U_2 = 0$ (from an outside option). The existing queueing literature typically assumes that he is perfectly rational: if $U_1 \geq 0$, he will join the system; otherwise, he will balk. However, accurately computing expected waiting and service time is typically not an easy task for customers. We thus depart from this literature by incorporating a more realistic assumption: customers lack the capability to accurately estimate their expected waiting time. As a result, customers cannot *guarantee* that the best choice is *always* chosen, and they may make mistakes. To formally model this noisy waiting time estimation, we introduce a random error term, ε, into customer's expected waiting time estimation. If $V_1 \equiv R - p - C(\mathbb{E}\,w + \varepsilon) \geq 0$, he will join the system and balk otherwise. As outside observers, we then obtain the customer joining probability $\varphi \equiv \mathbb{P}(\varepsilon \leq U_1/C)$, which should be interpreted as the *fraction* of customers who will join. For analytical tractability, we assume that the error term ε follows a logistic distribution $F(x) = 1/(1 + e^{-x/\theta})$ for some $\theta > 0$. The logistic distribution provides a good approximation to the normal distribution but has heavier tails (Talluri and van Ryzin 2004, pp. 305–306). Following McFadden (1974) and Anderson et al. (1992), we thus obtain the customer joining probability:

$$\varphi = \frac{e^{U_1/(C\theta)}}{1 + e^{U_1/(C\theta)}}.$$

It is important to note that in our model customers do *not* choose to play mixed strategies. It is the noisy estimation that drives this behavior, and φ denotes the fraction of customers that join the system.

To interpret the meaning of θ, note that the standard deviation of ε is $\sigma \equiv \sqrt{\mathrm{Var}(\varepsilon)} = (\pi/\sqrt{3})\theta \approx 1.8\theta$. Hence, the parameter θ is proportional to the standard deviation of the error term ε. Thus, the parameter θ measures the *error level* of customer expected waiting time estimation. (See Hey and Orme 1994, p. 1301, for a similar approach and explanations; alternatively, one could normalize the utility value while using the variance of the error term to capture the level of bounded rationality.) Furthermore, the standard deviation of customer expected utility V_1 is $\sigma_{V_1} \equiv C\sigma \approx 1.8C\theta$. For convenience, we define $\beta \equiv C\theta$, which measures the error level of customer expected utility estimation. This error level β reflects customer bounded rationality in the sense that customers have limited computational capability in perfectly estimating their expected waiting time (and as a result, their expected utility of joining) in the queueing setting. Hence, we interpret the parameter β as the extent customers are incapable of implementing the optimal decision because of their incapability in estimating expected waiting time. As $\beta \to 0$, the joining behavior converges to *full rationality*. At the other extreme, as $\beta \to \infty$, customers join or balk with equal fractions. Therefore, we can refer to the magnitude of β as the *level of bounded rationality*.

The interpretation of the level of bounded rationality β also follows from the well-known interpretation of the coefficients of logit regressions in that it captures the idea that better options are chosen more often. One can rewrite the joining probability as $\log(\varphi/(1-\varphi)) = (1/\beta)U_1$. The left-hand side is the "log odds" of joining the system, so β is the inverse of the difference in the log odds for any one-unit increase in the expected utility of joining the system. For example, when $\beta = 0.5$, then the log odds doubles for any one-unit increase in the expected utility of joining the system; when $\beta = 2$, then the log odds decreases by half for any one-unit increase in the expected utility of joining the system.

Given that β is the standard deviation of customer utility, it has the same unit as the reward R. We expect the magnitude of β to depend on the context. For example, McKelvey and Palfrey (1995) estimated $1/\beta$ in their game settings, which suggests that β ranges from 0.3 to 6.7 for a 1982 penny. Bajari and Hortacsu (2005, Table 7) reported the estimates for $1/\beta$: 15.68, 17.36, or 11.32 for a 1989 dollar. In another auction setting, Goeree et al. (2002) found out that the parameter β is around 0.09, 0.16, or 0.26 (see Table 4, p. 258) for a 2001 dollar. In the queueing setting, Kremer and Debo (2012) recently ran experiments in laboratories to test bounded rationality for visible queues. They found that β is statistically significantly different from zero: from their estimation, $1/\beta$ is 0.296, and the standard deviation is 0.025. They use the expected reward $R = \$10.2$.

Given that β has the same unit as R, one can use β/R to measure the level of bounded rationality relative to the reward. Note that this ratio is dimensionless and thus allows a unified comparison across studies without converting monetary units. Straightforward computation yields the magnitude of β/R in the literature: $\beta/R \in [0.06, 5.68]$ in McKelvey and Palfrey (1995), $[0.0038, 0.0059]$ in Bajari and Hortacsu (2005), $[0.018, 0.052]$ in Goeree et al. (2002), and around 0.33 in Kremer and Debo (2012). In our numerical study, β/R is in the range of $[0, 0.5]$ or $[0, 5]$ for the low-reward case and high-reward case, respectively, which is comparable to the magnitude observed in the literature.

We believe that people are typically not capable of accurately computing expected waiting time, and we interpret bounded rationality in terms of incapabilities in accurately estimating it in this paper. However, this is not the only possible interpretation. There are a couple of closely related interpretations in the literature that could be adopted here. First, Chen et al. (1997) propose the boundedly rational Nash equilibrium where agents are not utility maximizers but instead choose randomly in a fashion that is influenced by a *subconscious* utility function. In our setting, one can interpret U_1 as the subconscious utility because it is not explicitly known to the customer because of his bounded rationality. Moreover, Chen et al. (1997, p. 34) point out that "any mathematical structure commonly used with the noise or random utility interpretations can also be interpreted as a model of bounded rationality." Along this line, one can interpret our model in several ways by rewriting the customer utility function: $V_1 = R - p - C\,\mathbb{E}\,w + \varepsilon' = U_1 + \varepsilon'$. The noise term does not have to come from expected waiting time. It can stem from customer noisy perception of the waiting cost C or heterogeneity in customer preference for a particular queueing environment (see Maister 1985, Larson 1987): $\varepsilon' = -\varepsilon\,\mathbb{E}\,w$.

The second different interpretation comes from Mattsson and Weibull (2002): that agents are utility maximizers. However, they have to make some effort in order to implement any desired outcome, and the disutility of this effort enters their utility function. In our setting, one may think of customers having to make some effort to implement the optimal joining decision, perhaps because of the challenging task of estimating the expected waiting time. These different interpretations provide different justifications to our model. We follow the interpretation of incapability of accurately estimating the expected waiting time because we believe that it captures best the main computational limitation customers face in service systems.

Our model has both usefulness and limitations. The model is useful because (i) it provides a systematic way to capture bounded rationality in service systems using a single parameter β; (ii) this single parameter β is endowed with a concrete meaning that includes both conventional full rationality and purely random behavior (i.e., full bounded rationality), and it allows us to investigate the impact of bounded rationality (e.g., in terms of revenue, welfare, and pricing); (iii) this model is flexible to accommodate several interpretations; and (iv) it can be readily used for further empirical or experimental tests because of its close connection to commonly used logit regressions. However, this model has limitations: flexibility comes with both advantages and disadvantages. That the model itself cannot be pinned down to a unique interpretation without lab experiments calls for further empirical or experimental studies. Indeed, that is one of the goals of the paper: to stimulate a desirable theoretical–empirical feedback loop in service operations management research. Also, the logit model is a special case of the general class of the "Fechner" approach of modeling the stochastic element in decision making, and we refer the reader to Loomes et al. (2002) for other models that can potentially be useful.

Both the expected waiting and service time $\mathbb{E}\,w$ and the level of bounded rationality depend on the configuration of the service system: whether the queue length is observable to customers or not. Hereafter, we distinguish the visible and invisible queues, and denote the levels of bounded rationality β and β_I for them, respectively.

2.1. The Visible Queue

Consider a single-server queueing system that is observable. Customers arrive to the system according to a Poisson process with rate λ. Upon arrival, each customer decides whether or not to join the queue after observing the queue length and based on his estimate of the waiting time. Service times are assumed to be independently, identically, and

exponentially distributed with mean $1/\mu$. Denote the utilization of the system $\rho \equiv \lambda/\mu$ if all customers join. Customers are served on a first-come, first-served basis. Upon arrival, observing n customers in the system, the fraction of customers that join the system is given by

$$\varphi_n \equiv \frac{e^{(R-p-((n+1)C)/\mu)/\beta}}{1+e^{(R-p-((n+1)C)/\mu)/\beta}}, \tag{1}$$

for $n = 0, 1, 2, \ldots$. To compute the expected waiting time, each customer needs perfect information when he joins about the number of customers and the service rate and needs the cognitive capability to transform this information into an estimate of the expected waiting time. Thus, if customers lack either the information or the cognitive capability, then the waiting time estimate will be inaccurate. Further, the accuracy of the expected waiting time is severely impacted because they need to do this in real time.

For ease of exposition, we let $\lambda_n \equiv \lambda\varphi_n$, $n = 0, 1, 2, \ldots$, be the state-dependent queue-joining rates. Then, we can treat the number of customers in the system as a birth–death process with birth rate λ_n and death rate μ. Although customers are boundedly rational, we first show that the *stability* of the system is guaranteed as long as β is finite, as stated in the following proposition.

Proposition 1. *The visible queueing system with boundedly rational customers is stable for $\beta < \infty$, and the probability distribution in steady state is as follows*:

$$P_0 = \frac{1}{1+\sum_{k=1}^{\infty}(\lambda_0\lambda_1\cdots\lambda_{k-1})/\mu^k}$$

is the probability that the system is in state 0, *and*

$$P_n = \frac{\lambda_0\lambda_1\cdots\lambda_{n-1}}{\mu^n(1+\sum_{k=1}^{\infty}(\lambda_0\lambda_1\cdots\lambda_{k-1})/\mu^k)}$$

is the probability in state n, $n \geq 1$.

Note that the queue length distribution becomes unbounded from above for $\beta > 0$, but the system is always stable by Proposition 1.

We numerically observe that both utilization (i.e., $\rho(p,\beta) \equiv \sum_{n=0}^{\infty}\lambda_n P_n/\mu$) and expected queue length (i.e., $q(p,\beta) \equiv \sum_{n=0}^{\infty} nP_n$) have an intricate relationship with the level of bounded rationality β. In particular, neither of them is monotonic or unimodal as a function of the level of bounded rationality β. To understand why this happens, we first define $n_s = [((R-p)\mu)/C]$ as the threshold queue length used by fully rational customers in deciding to join the queue or not. We then divide the states of the system into two regions using the threshold n_s: Region 1 comprises the states when the number of customers in the system is less than n_s; Region 2 comprises all the other states, i.e., those when the number of customers in the system is greater than n_s. When customers are fully rational, customers will join with probability 1 in Region 1 and 0 in Region 2. For every strictly positive β, the joining probability will be between 0 and 1. Hence, bounded rationality in Region 1 lowers utilization (and expected queue length), and it increases utilization (and expected queue length) in Region 2. As customers become more boundedly rational, these two effects take place simultaneously. It turns out that it is not clear which effect dominates.

2.2. The Invisible Queue

We now turn to an invisible queueing system using the same model setup as §2.1. The only difference is that the queue length is *invisible* to customers. Potential customers arrive to this system according to a Poisson process with rate λ. Because customers cannot observe the state of the system, they have to make a decision a priori whether to arrive to the queue or not. Different from the visible-queue setting, each customer has to form beliefs about the state of the system, which is determined by other customers' strategies. We assume that each customer knows all the underlying parameters of the system such as R, C, p, λ, and μ. He knows that he is boundedly rational, and all the other customers are also boundedly rational. In other words, he is able to form the correct belief about the state of the system, which is used to estimate his expected waiting time.

In investigating the system, we are initially interested in the fraction of customers that join the system $\varphi(p,\beta_I) \in [0,1]$ in equilibrium. Again, customers do not choose to play mixed strategies. Only from the point of view of an outside observer are customer decisions probabilistic. A customer's net benefit or utility of joining is $U_1 = R - p - C\mathbb{E}\,w = R - p - C/(\mu - \varphi(p,\beta_I)\lambda)^+$, where we used the fact that the *thinning* of a Poisson process with arrival rate λ is still a Poisson process with rate $\varphi(p,\beta_I)\lambda$ and $C/(\mu-\varphi(p,\beta_I)\lambda)^+$ is the customer's expected waiting cost. The arrival rate $\varphi(p,\beta_I)\lambda$ of the queue will be referred to as *effective* demand. According to our model of bounded rationality in §2, each customer cannot perfectly estimate his expected waiting time and hence joins the system with probability $\varphi_I \equiv e^{U_1/\beta_I}/(1+e^{U_1/\beta_I})$. In equilibrium, *consistency* requires that the system effective arrival rate is consistent with customer behavior: $\varphi(p,\beta_I) = \varphi_I$. Hence, we define the equilibrium of the invisible queueing system as follows.

Definition 1 (Equilibrium Joining Fraction). We say that $\varphi(p,\beta_I)$ is an equilibrium joining fraction if it satisfies the following:

$$\varphi(p,\beta_I) = \frac{e^{(R-p-C/(\mu-\varphi(p,\beta_I)\lambda)^+)/\beta_I}}{1+e^{(R-p-C/(\mu-\varphi(p,\beta_I)\lambda)^+)/\beta_I}}, \tag{2}$$

for $\beta_I > 0$, and

$$\varphi(p, 0) = \min\{\varphi_0, 1\},$$

where φ_0 satisfies

$$R - p - \frac{C}{\mu - \varphi_0 \lambda} = 0, \tag{3}$$

for $\beta_I = 0$.

When $\beta_I > 0$, Equation (2) yields a fixed-point problem given that the logit expression in the right-hand side includes the equilibrium joining fraction (i.e., the left-hand side).

When $\beta_I = 0$, i.e., customers are fully rational, then the definition is precisely Hassin's (1986) equilibrium condition (Equation (4.1), p. 1189). It is possible that there is no $\varphi_0 \in [0, 1]$ satisfying Equation (3), and the actual arrival rate then is λ because even if all customers decide to join, each customer's expected utility is still strictly positive. According to this definition, we have $\varphi(p, 0) = \min\{\mu/\lambda - C/(\lambda(R - p)), 1\}$.

The assumption that customers are boundedly rational in their strategies but not in beliefs is consistent with the economics literature of modeling bounded rationality (see Chen et al. 1997 and references therein). This approach is certainly restrictive because decision makers are capable of correctly calculating the expected error-prone actions of the other players, which are certainly nontrivial cognitive tasks (Mallard 2011). But McKelvey and Palfrey (1995) and Chen et al. (1997) show that such quantal response equilibria emerge from learning. Hence, in the short-run or transient states, customers may neglect others' bounded rationality, but eventually they would be able to form the correct belief about others' strategies so that the fixed-point outcome according to Definition 1 would emerge. We relax Definition 1 in Appendix D in our technical report (Huang et al. 2012) by allowing customers to have incorrect beliefs. There, the model is closely related to the "level-k thinking" (Stahl and Wilson 1995): the closed-loop, fixed-point type of Equation (2) would become open-loop. The resulting analysis for both revenue and social welfare maximization is straightforward. We refer the reader to Appendix D of Huang et al. (2012) for detailed discussion. In §2 of the online supplement, we prove that for a given range of β_I, the path of customer joining decisions over time, when the customers adaptively learn the expected waiting time, converges to the equilibrium in Definition 1. Also, focusing on bounded rationality as incapable of accurately predicting the expected waiting time allows a fair comparison of the visible queue versus invisible queue.

Next, we investigate whether an equilibrium always exists. Proposition 2 shows that there always exists a *unique* equilibrium.

Proposition 2. *There always exists a unique equilibrium for the invisible queue for any finite price p and level of bounded rationality $\beta_I > 0$.*

We are now interested in how the (unique) equilibrium $\varphi(p, \beta_I)$ behaves as a function of the price p and the level of bounded rationality β_I. For convenience, we let $\bar{p} \equiv R - (2C)/(2\mu - \lambda)$ denote the price under which each customer receives exactly zero utility so that the equilibrium joining fraction is half *regardless of* the level of bounded rationality. The following proposition characterizes the equilibrium joining fraction.

Proposition 3. (i) *If $p < \bar{p}$, equilibrium joining fraction $\varphi(p, \beta_I)$ is strictly decreasing in β_I.*

(ii) *If $p > \bar{p}$, equilibrium joining fraction $\varphi(p, \beta_I)$ is strictly increasing in β_I.*

(iii) *If $p = \bar{p}$, equilibrium joining fraction $\varphi(p, \beta_I) = 1/2$ for any β_I.*

(iv) *For any fixed β_I, equilibrium joining fraction $\varphi(p, \beta_I)$ is strictly decreasing in p.*

We offer the following intuition: when the price is so low that each customer receives strictly positive utility, the initial joining fraction is above half. As the level of bounded rationality increases, better decisions are made less often, and thus the joining fraction decreases as customers are more boundedly rational. Interestingly, if the price is set so that each customer receives exactly zero utility in equilibrium, then increasing the level of bounded rationality has no effect on the joining fraction because customers join or balk with equal fractions regardless of the level of bounded rationality.

It is intuitively clear that $\varphi(p, \beta_I)$ is strictly decreasing in price p by equality (2); i.e., a larger price always results in a lower joining fraction, regardless of the level of bounded rationality, which is the "law of demand" in this service setting.

It is useful to note that the invisible queue with boundedly rational customers is essentially an $M/M/1$ system with arrival rate $\varphi(p, \beta_I)\lambda$ and service rate μ. The server utilization is $\rho_I(p, \beta_I) \equiv \rho\varphi(p, \beta_I)$. Thus the utilization behaves the same as the joining fraction as a function of p and β_I (which is already characterized in Proposition 3). Using similar logic, we can characterize the expected queue length $q_I(p, \beta_I) \equiv \varphi(p, \beta_I)\lambda/(\mu - \varphi(p, \beta_I)\lambda)$ as a function of p and β_I.

3. Revenue Maximization

In the previous section, our attention was focused on the system equilibrium or dynamics. In this section, we focus our attention on the revenue generated from such systems. In this sense, we are looking from a revenue-maximizing firm's perspective.

3.1. The Visible Queue

The revenue as a function of price p and level of bounded rationality $\beta > 0$ is

$$\Pi(p, \beta) \equiv \sum_{n=0}^{\infty} \lambda_n P_n p = \sum_{n=0}^{\infty} \frac{e^{(R-p-((n+1)C)/\mu)/\beta}}{1 + e^{(R-p-((n+1)C)/\mu)/\beta}} \lambda P_n p.$$

Note that we normalize the cost of serving customers to zero without loss of generality.

When customers are fully rational, i.e., $\beta = 0$, we naturally define $\Pi(p, 0) \equiv \lim_{\beta \to 0} \Pi(p, \beta)$ for any price p (one can show that such a limit exists). In the setting with fully rational customers, Naor (1969) shows that choosing the revenue-maximizing price boils down to choosing the optimal integer n to maximize the revenue function $\Pi_n = \lambda((1-\rho^n)/(1-\rho^{n+1}))(R - Cn/\mu)$ so that $p(n) = R - Cn/\mu$. Let n_r be the maximizer and Π_{n_r} be the maximized revenue.

We are interested in comparing the optimal revenue $\Pi(p^*(\beta), \beta)$ when the revenue-maximizing price $p^*(\beta)$ is set, if customers are slightly boundedly rational, and the optimal revenue $\Pi_{n_r} \equiv \sup_p \lim_{\beta \to 0} \Pi(p, \beta)$ if customers are fully rational. For convenience, let $p^* \equiv p^*(0) = R - Cn_r/\mu$ be the revenue-maximizing price under full rationality.

Proposition 4. *$p^*(\beta) < p^*$ and $\Pi(p^*(\beta), \beta) < \Pi_{n_r}$ when β is strictly positive but sufficiently small.*

Proposition 4 does not extend to situations when the level of bounded rationality becomes high. (As customers become fully boundedly rational, the optimal revenue goes to infinity.) The intuition behind Proposition 4 is as follows: a strictly positive and small β forces the revenue-maximizing firm to strictly lower its price compared to full rationality, which in turn brings strictly lower revenue for the firm. In other words, a strictly positive and small β makes revenue collecting less profitable. The reason is that a strictly positive and small β strictly reduces the effective demand of the system, ceteris paribus. A strictly lower price is necessary to increase the effective demand to maximize revenue.

From our numerical studies, we find that the optimal revenue is not necessarily monotone with respect to the level of bounded rationality, and the revenue-maximizing price as a function of the level of bounded rationality can increase or decrease. The explanation is similar to why the utilization is not necessarily monotonic in the level of bounded rationality because the revenue directly depends on the utilization: $\Pi(p, \beta) = \rho(p, \beta)\mu p$.

Impact of Ignoring Bounded Rationality. Without taking into account customer bounded rationality, the revenue-maximizing firm will rationally charge price $p^*(0) = R - Cn_r/\mu$. We are interested in the revenue loss as a result of bounded rationality and carried out a numerical study. We approximate the steady-state distribution by truncating the birth–death process to a finite state space, gradually increasing the number of states until the revenue function is no longer sensitive to the truncation level. To investigate the impact of the level of bounded rationality β, the reward-to-cost ratio R/C, and the traffic intensity λ/μ on the revenue loss, we normalize $C = 1$ and $\mu = 1$ and vary R and λ. Our systematic numerical study is intended to include high/low utilization crossed with high/low reward. Figure 1 shows a representative example where $R \in \{2, 20\}$ and $\lambda \in \{0.5, 5\}$. From Figure 1, we have the following observations: (i) The revenue loss can be significant (more than 200%, for instance), depending on the parameters, and can be arbitrarily large as β goes to infinity. (ii) The revenue loss is not necessarily monotone with respect to β. This is likely because the revenue-maximizing price is not necessarily monotone in β. To understand this fact, recall that neither the utilization nor expected queue length are necessarily monotone, and our explanation in §2.1 applies here given that these basic measures drive the revenue.

3.2. The Invisible Queue

The firm's objective is to choose a price p to maximize the expected revenue $\Pi^I(p, \beta_I) \equiv p\varphi(p, \beta_I)\lambda$, where $\varphi(p, \beta_I)\lambda$ is the effective demand rate to the system.

To investigate the firm's revenue maximization problem, we first study the behavior of the revenue $\Pi^I(p, \beta_I)$ as a function of price p and level of bounded rationality β_I. Note that $\Pi^I(p, \beta_I)$ is simply a linear transformation of $\varphi(p, \beta_I)$; hence, Proposition 3 characterizes how $\Pi^I(p, \beta_I)$ behaves as a function of β_I for any fixed p.

We next investigate how revenue $\Pi^I(p, \beta_I)$ behaves as a function of price p for any fixed level of bounded rationality β_I. To state Proposition 5, we denote $\beta_0 \equiv R/2 - 2C\mu/(2\mu - \lambda)^2$, which is the level of bounded rationality at which the optimal price $p^*(\beta_0) = \bar{p}$. Hence, at the level of bounded rationality β_0, each customer receives zero expected utility of joining under the revenue-maximizing price.

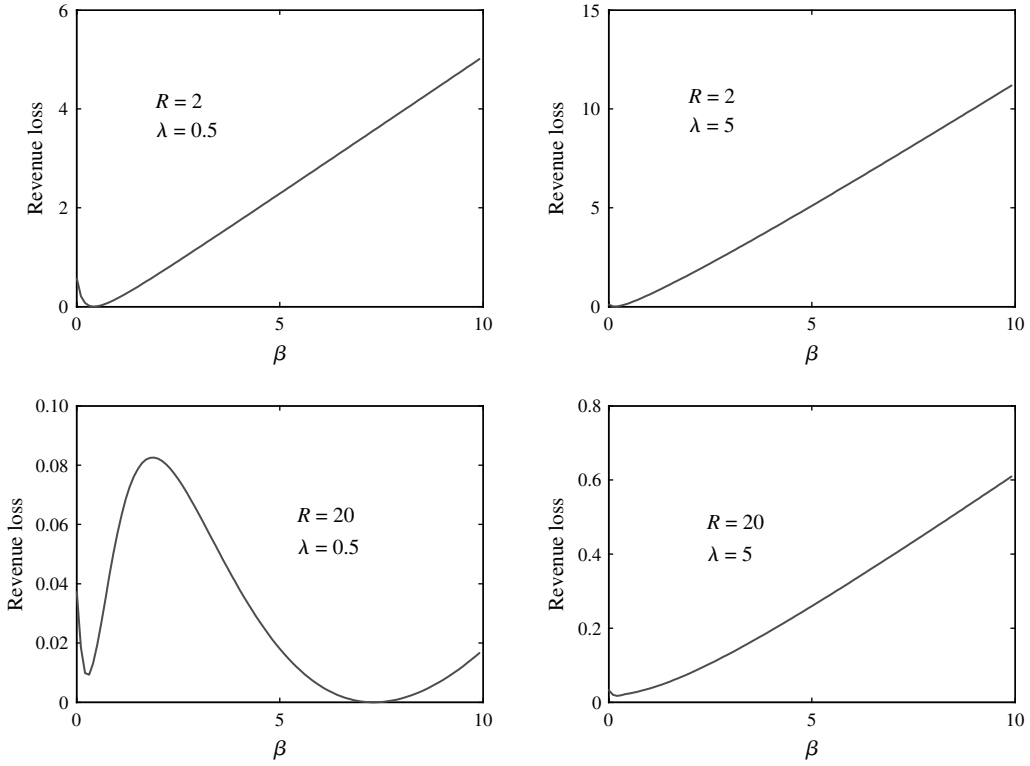
Proposition 5. (i) *For any fixed level of bounded rationality β_I, $\Pi^I(p, \beta_I)$ is unimodal in p, and thus there exists a unique price $p^*(\beta_I)$ that maximizes $\Pi^I(p, \beta_I)$.*

(ii) *The optimal price $p^*(\beta_I)$ is strictly increasing in the level of bounded rationality β_I for $\beta_I \in [\max\{\beta_0, 0\}, \infty)$.*

From this proposition, we obtain that the revenue-maximizing price $p^*(\beta_I)$ is monotonically increasing in $\beta_I \in [0, \infty)$ if R is sufficiently small.

When the optimal price induces each customer to receive strictly negative expected utility in equilibrium, a higher β_I would induce the firm to increase its price. The reason is that higher β_I leads to higher

Figure 1 **Revenue Loss When Revealing the Queue if Bounded Rationality Is Ignored ($C = 1, \mu = 1$, Revenue Loss $\Delta\Pi(\beta) \equiv (\Pi(p^*(\beta), \beta) - \Pi(p^*(0), \beta))/\Pi(p^*(0), \beta)$)**



joining fractions for a fixed price in this case, according to Proposition 3(ii). Hence, when the optimal price is sufficiently high (so that each customer receives strictly negative expected utility in equilibrium), then increasing β_I leads to even higher optimal prices. However, when the optimal price is low so that each customer receives strictly positive utility, then increasing β_I *can* lead to lower optimal prices, where the firm's trade-off is about the benefit of higher prices versus the loss of lower effective demand. Proposition 5 above partially characterizes which one of these effects dominates.

We are now ready to state the result on the effect of the level of bounded rationality on the *optimal* revenue $\Pi^I(p^*(\beta_I), \beta_I)$. Using the envelope theorem, we obtain the following immediate corollary to Proposition 3.

Corollary 1. (i) *If* $p^*(\beta_0) > \bar{p}$, *then*

$$d\Pi^I(p^*(\beta_I), \beta_I)/d\beta_I\big|_{\beta_I=\beta_0} > 0.$$

(ii) *If* $p^*(\beta_0) < \bar{p}$, *then* $d\Pi^I(p^*(\beta_I), \beta_I)/d\beta_I|_{\beta_I=\beta_0} < 0$.
(iii) *If* $p^*(\beta_0) = \bar{p}$, *then* $d\Pi^I(p^*(\beta_I), \beta_I)/d\beta_I|_{\beta_I=\beta_0} = 0$.

By Proposition 5 and Corollary 1, we know that the optimal revenue $\Pi^I(p^*(\beta_I), \beta_I)$ strictly increases in β_I as β_I is sufficiently large. Therefore, the revenue-maximizing firm can exploit the bounded rationality when β_I is sufficiently large.

Finally, we are also interested in how the arrival rate affects revenue because it would later be useful. Recall that in Hassin (1986) where customers are fully rational, there exists some λ_0, when $\lambda > \lambda_0$, the revenue function is independent of λ. Interestingly, in our case with boundedly rational customers, we have that a higher arrival rate λ always leads to *strictly* higher revenue.

Proposition 6. *For any fixed price* p *and level of bounded rationality* $\beta_I > 0$, *the equilibrium joining fraction is strictly decreasing and the revenue is strictly increasing in the arrival rate* λ.

The result that higher arrival rates lead to lower equilibrium joining fractions is not surprising because more congestion forces each customer to lower his joining probability. However, the result that more arrivals always lead to more revenue may appear to be surprising. The key insight is that the marginal revenue increment has to be proportional to the marginal joining fraction decrement given the equilibrium condition (2). Hence, the effective demand $\varphi(p, \beta_I)\lambda$ increases in λ. Proposition 6 implies that for any price p, not necessarily the optimal price, higher arrival rates lead to higher revenue. In particular, higher arrival rates lead to higher *optimal* revenue. Such a finding is in stark contrast to Hassin's (1986) full rationality case.

Impact of Ignoring Bounded Rationality. Finally, we are interested in the consequence of ignoring bounded rationality while customers are actually boundedly rational. Without taking into account customer bounded rationality, the revenue-maximizing firm will rationally charge price $p^*(0)$, which is generally different from the revenue-maximizing price $p^*(\beta_I)$. Hence, $\Pi^I(p^*(0), \beta_I) \equiv p^*(0)\varphi(p^*(0), \beta_I)\lambda \le \Pi^I(p^*(\beta_I), \beta_I)$. We are interested in the revenue loss $\Delta\Pi^I(\beta_I) \equiv (\Pi^I(p^*(\beta_I), \beta_I) - \Pi^I(p^*(0), \beta_I))/\Pi^I(p^*(0), \beta_I)$ as a result of this ignorance of bounded rationality. As an example, we use the same parameters as before. Figure 2 shows that the revenue loss can be nontrivial (e.g., more than 200%). In general, the revenue loss is not necessarily monotone with respect to β_I. Similar to the visible queue, this observation could be driven by the fact that the revenue-maximizing price is not necessarily monotone in β_I (Proposition 5). Hence, when β_I increases, the revenue-maximizing price $p^*(\beta_I)$ and $p^*(0)$ may become closer, which would result in lower revenue loss; the case when $R = 20$, $\lambda = 0.5$, and $\beta_I = 5$ in Figure 2 illustrates this point. However, because β_I is larger than a certain threshold, the loss is significant. In fact, $\lim_{\beta_I \to \infty} \Delta\Pi^I(\beta_I) = \infty$; i.e., the revenue loss can be arbitrarily large as β_I is sufficiently large. This directly follows from the fact that $\lim_{\beta_I \to \infty} p^*(\beta_I) = \infty$ (see the first-order condition in the proof of Proposition 5) and $\lim_{\beta_I \to \infty} \varphi(p, \beta_I) = 0.5$.

4. Social Welfare Maximization

We now turn to study the problem from a social planner's perspective. The social planner is interested in maximizing social welfare. In this section, we study the impact of bounded rationality on the social welfare, both when the price is exogenously given and when the social planner charges the welfare-maximizing price.

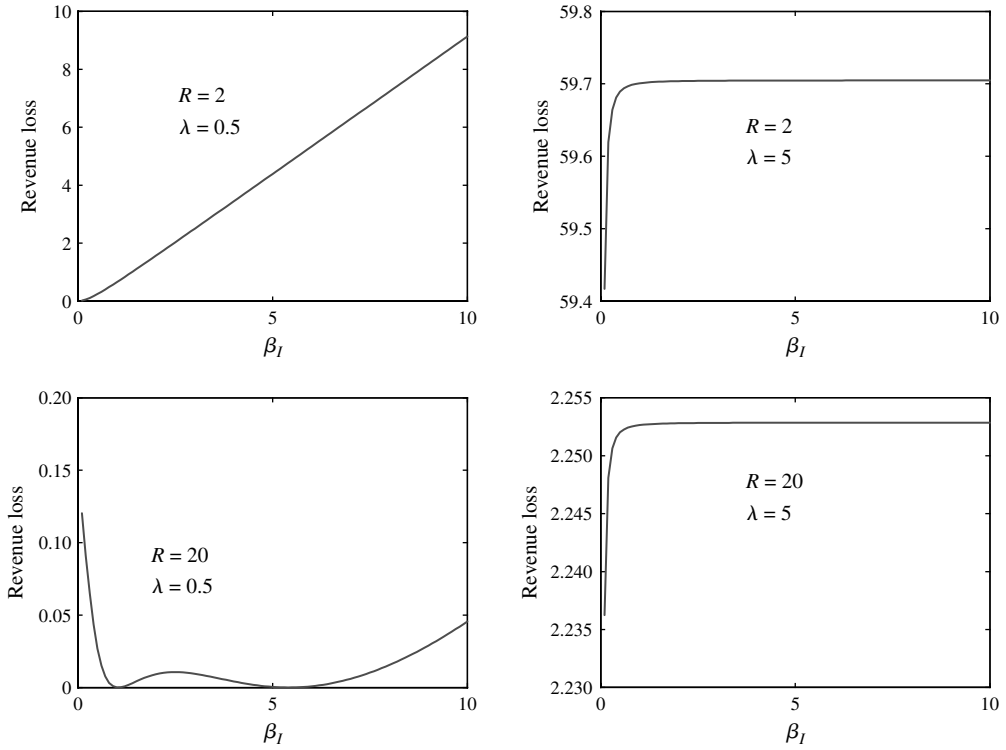
4.1. The Visible Queue

In many settings, the price is set or influenced by other considerations such as market conditions, competition, or a price being set by a third party. There are settings where optimizing over the price or even charging a price may not be feasible—for example, toll charges on public roads, food distribution in natural disasters, etc. We first study how bounded rationality affects social welfare for a given price. Observe that the fixed price p always appears as $R - p$ in Equation (1). For the ease of comparing with the results in the classic paper by Naor (1969) and for mathematical convenience, we assume $p = 0$. However, the findings extend to the setting where the price is nonzero. We can derive the social welfare function as follows:

$$W(\beta) \equiv W(p, \beta)|_{p=0} = \sum_{n=0}^{\infty} \lambda_n P_n R - \sum_{n=0}^{\infty} n P_n C. \quad (4)$$

The first term in Equation (4) is the (long-run) average reward, and the second term is the average waiting

Figure 2 **Revenue Loss When Hiding the Queue if Bounded Rationality Is Ignored ($C = 1, \mu = 1$, Revenue Loss $\Delta\Pi^I(\beta_I) \equiv (\Pi^I(p^*(\beta_I), \beta_I) - \Pi^I(p^*(0), \beta_I))/\Pi^I(p^*(0), \beta_I)$)**



cost. Notice that if customers are fully rational, i.e., $\beta = 0$, then $P_n = 1$ for $n = 0, 1, \ldots, [R\mu/C] - 1$ and $P_n = 0$ for $n > [R\mu/C] - 1$, in which case our model reduces to Naor's (1969) model.

To compare social welfare $W(\beta)$ with $W(0)$, we first recall $n_s = [R\mu/C]$ as the threshold queue length used by fully rational customers in deciding to join the queue or not and n_0 as the equivalent threshold from a social planner's point of view. Naor (1969) shows that $n_s \geq n_0$; i.e., self-interested customers typically make the system more congested than the socially optimal level.

Intuitively, bounded rationality can create two effects for the social welfare: a positive (i.e., welfare-improving) effect and negative (i.e., welfare-diminishing) effect. To understand how these two effects come into play, we divide the states of the system into three regions using the two thresholds n_s and n_0: Region 1 comprises the states when the number of customers in the system is less than n_0; Region 2 comprises the states when the number of customers in the system is greater than n_0 but less than n_s; and Region 3 comprises all the other states (i.e., those when the number of customers in the system is greater than n_s). When customers are fully rational, customers will join with probability 1 in Regions 1 and 2 and 0 in Region 3. Recall that to maximize social welfare, customers should join with probability 1 in Region 1 and 0 in Regions 2 and 3. However, for any strictly positive level of bounded rationality, the joining fraction is between 0 and 1. Hence, social welfare will decrease in Regions 1 and 3 but increase in Region 2 compared to full rationality. As customers become more boundedly rational, these effects take place simultaneously, and it seems unclear a priori which effect would dominate.

Although Equation (4) presents a complete characterization of the social welfare in terms of the level of bounded rationality β, the dependence is quite intricate. Thus, we begin by analyzing the social welfare $W(\beta)$ in the neighborhood of zero. We are interested in the relationship between $W(\beta)$ and $W(0)$ when β is sufficiently small. It turns out we are able to completely characterize the conditions in which one effect dominates the other. We have the following simple inequalities showing when the social welfare increases or decreases as the customers become slightly boundedly rational.

Proposition 7. *If any one of the following three conditions is satisfied,*

(1) $n_s < R\mu/C - 1/2$,
(2) $n_s = n_0$, *and*
(3) $n_s = R\mu/C - 1/2$ *and* $\rho > 1$,

then $W(\beta) < W(0)$ *when* $\beta > 0$ *is sufficiently small. Otherwise,* $W(\beta) > W(0)$ *when* $\beta > 0$ *is sufficiently small.*

According to Proposition 7, if either of the following two conditions is satisfied,

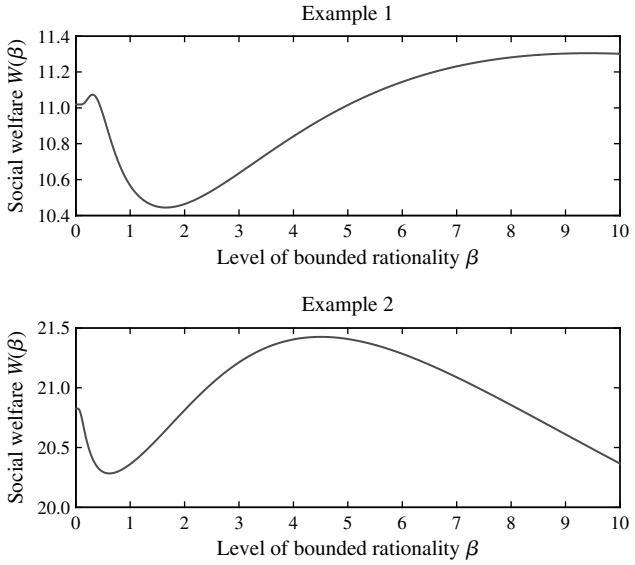
(a) $n_s > n_0$, and $n_s > R\mu/C - 1/2$, and
(b) $n_s > n_0$, $n_s = R\mu/C - 1/2$ and $\rho \leq 1$,

then a strictly positive and small β *strictly improves* the social welfare. Finding a simple sufficient and necessary condition for $n_s > n_0$ is difficult. However, Lemma EC.5 in the appendix of Huang et al. (2012) shows that either of the following two conditions is sufficient for $n_s > n_0$: (1) $\rho > 1$ and $n_s > 1$ and (2) $\sqrt{2} - 1 < \rho < 1$ and $n_s > 2$.

In other words, compared with fully rational customers, boundedly rational customers err on both sides, joining a more congested system and balking when congestion is low. Although the former is detrimental to the social welfare, the latter can be beneficial. The social welfare can thus be improved depending on which of these effects dominates. In Proposition 7, we provide a simple characterization of this dichotomy. This result appears to be striking: although bounded rationality is usually associated with suboptimal decisions, it might yield better outcomes for the society overall. This is due to the externality present among the boundedly rational customers in the system.

As we discussed before, characterizing the social welfare as a function of the level of bounded rationality is difficult because of the intricate joint effects coming from the three regions simultaneously as customers become more boundedly rational. To understand the social welfare as a function of the level of bounded rationality, we carry out a numerical study. To demonstrate that the social welfare function is not necessarily unimodal in a reasonable range of bounded rationality, we intentionally use different sets of parameters. In the first example, the parameters are $R = 14.93, C = 7, \mu = 3,$ and $\lambda = 5$, so that $n_s = 6 > R\mu/C - 1/2 = 5.8986$. As shown in the graph in the upper panel of Figure 3, the social welfare strictly increases initially as predicted by Proposition 7; however, it decreases and then increases again when the customers become more boundedly rational. In the second example, the parameters are $R = 16, C = 7, \mu = 3,$ and $\lambda = 2.6$, so that $v_s = 6.8571$ and $n_s = 6 < R\mu/C - 1/2 = 6.3571$. As illustrated in the graph in the lower panel of Figure 3, the social welfare initially decreases as predicted by Proposition 7; however, it increases as the level of bounded rationality becomes larger and decreases again as the level of bounded rationality further increases. Thus, even though the social welfare is well behaved for a strictly positive and small β, it does not possess global properties such as convexity/concavity or even unimodality. This result stands in contrast to the invisible queue where the social welfare function is unimodal in the level of bounded rationality.

Figure 3 **Global Behavior of Social Welfare: Example 1 ($R = 14.93$, $C = 7$, $\mu = 3$, $\lambda = 5$, $n_s = 6 > (R\mu)/C - 1/2 = 5.8986$) and Example 2 ($R = 16$, $C = 7$, $\mu = 3$, $\lambda = 2.6$, $v_s = 6.8571$, $n_s = 6 < (R\mu)/C - 1/2 = 6.3571$)**



We have analyzed the impact of bounded rationality on social welfare for a given price. However, the social planner may be able to freely charge a price to maximize the social welfare. We now investigate the implication of bounded rationality for the social welfare if the social planner can regulate the system by pricing optimally. We are interested in whether bounded rationality increases or decreases the social welfare. We denote the social welfare function $W(p, \beta)$ when the social planner charges price p and the customers' level of bounded rationality is β. Obviously, the social welfare $W(p, \beta)$ can be expressed in a similar fashion as Equation (4).

Naor (1969) shows that by levying tolls the social planner can achieve the social optimum when customers are fully rational. In particular, if any price $p^* \in (R - C(n_0 + 1)/\mu, R - Cn_0/\mu]$ is charged by the social planner, then the maximum social welfare $W^*(0) \equiv \sup_p W(p, 0)$ can be achieved. We study whether the optimal social welfare $W^*(0)$ can be achieved by adding bounded rationality on the part of customers.

We show that when facing boundedly rational customers, the first-best social welfare can never be achieved, as stated in the following proposition.

PROPOSITION 8. *For any price $p \in \mathbb{R}$ charged to customers, the social welfare $W(p, \beta)$ is strictly lower than the social optimum when β is strictly positive; i.e., $W(p, \beta) < W^*(0)$ for $\beta > 0$.*

This proposition proves that bounded rationality always results in social welfare losses compared with the full rationality case. This is in contrast to (a) Naor (1969), where levying tolls achieves the socially optimal welfare; (b) the result in Proposition 7 that a strictly positive and small β can increase the social welfare when an arbitrary price is charged (when the firm charges the optimal price, then only case (2) in Proposition 7 arises); and (c) the result in Proposition 10 of the invisible queue where there may not be any welfare loss due to bounded rationality. This stems from the following: when customers are fully rational, the social planner can always regulate the service system by charging prices to achieve the social optimality $W^*(0)$. However, each boundedly rational customer randomizes with nondegenerate probabilities to join or balk. In this case, the social planner loses the *precise* control over the customers' joining decisions, and thus bounded rationality dilutes the effectiveness of the price regulation. Of course, if the firm can implement state-dependent pricing, then the social planner can achieve the first-best, even in the presence of bounded rationality.

From our numerical studies, we find that there can be multiple prices that maximize the welfare for a given level of bounded rationality. Second, the optimal welfare is not necessarily monotone with respect to the level of bounded rationality. The explanation for the nonmonotonicity behavior is similar to that for the global nonmonotonicity behavior of the social welfare function with respect to β: bounded rationality has the positive and negative effect on welfare simultaneously.

Impact of Ignoring Bounded Rationality. Without taking into account customer bounded rationality, the social planner will pick one price in the range $(R - C(n_0 + 1)/\mu, R - Cn_0/\mu]$. We are interested in the welfare loss that results from bounded rationality. Using the same example as in the revenue maximization case, Figure 4 shows the welfare loss when the firm charges price $R - Cn_0/\mu$. Again, we observe nontrivial welfare loss (more than 60% for instance). The intuition from the nonmonotonicity behavior is similar to the explanation we provided for the global nonmonotonicity behavior of the social welfare function when $p = 0$: as β increases, the positive and negative effect on welfare occurs simultaneously.

4.2. The Invisible Queue

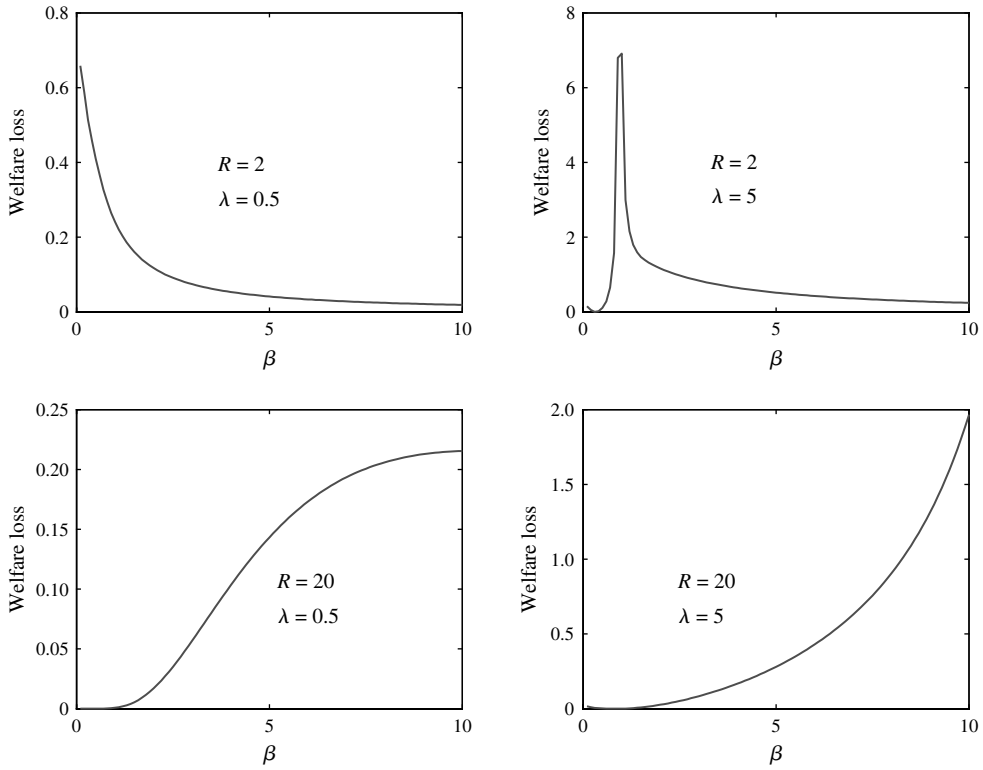
For any price p and level of bounded rationality β_I, the social welfare function is denoted as

$$W^I(\varphi(p, \beta_I)) \equiv \varphi(p, \beta_I)\lambda R - \frac{\varphi(p, \beta_I)\lambda}{\mu - \varphi(p, \beta_I)\lambda} C. \qquad (5)$$

For mathematical convenience, we may drop the dependence over $\varphi(p, \beta_I)$ and write $W^I(p, \beta_I)$.

The first term of Equation (5) is the average benefit the customers receive from the system, and the second term is the average waiting cost incurred by the customers. Note that price p affects social welfare

Figure 4 **Welfare Loss When Revealing the Queue if Bounded Rationality Is Ignored ($C = 1, \mu = 1$, Welfare Loss $\Delta W(\beta) \equiv (W(\max\{0, p_w^*(\beta)\}, \beta) - W(p^*(0), \beta))/(W(p^*(0), \beta))$)**



only indirectly through the equilibrium joining fraction $\varphi(p, \beta_I)$.

First, observe that the social welfare $W^I(\varphi(p, \beta_I))$ is strictly concave in $\varphi(p, \beta_I)$ (Lemma EC.2 in the appendix of Huang et al. 2012). Combining this fact with the characterization of the equilibrium joining fraction $\varphi(p, \beta_I)$, we can characterize how the social welfare behaves as a function of the level of bounded rationality in Proposition 9.

From Proposition 5, one can obtain that $p^*(0) = R(1 - \sqrt{C/(\mu R)})$. Note that when customers are fully rational, the welfare-maximizing price and the revenue-maximizing price coincide.

Proposition 9. (i) *If $p = \bar{p}$, then social welfare $W^I(\varphi(p, \beta_I))$ is constant for $\beta_I \geq 0$.*

(ii) *If $[p^*(0) - \bar{p}][p - \bar{p}] \leq 0$ and $p \neq \bar{p}$, then social welfare $W^I(\varphi(p, \beta_I))$ strictly increases for $\beta_I \geq 0$.*

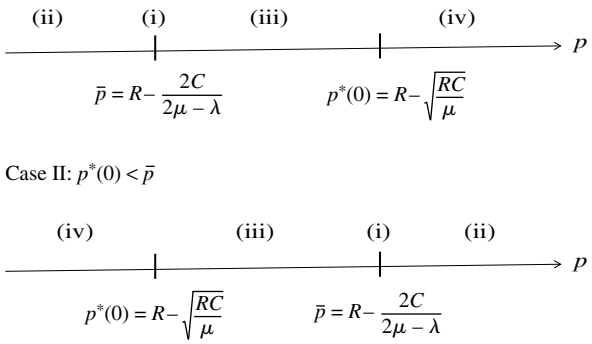
(iii) *If $p \in (\min\{p^*(0), \bar{p}\}, \max\{p^*(0), \bar{p}\}) \cup \{p^*(0)\}$, and $p^*(0) \neq \bar{p}$, then social welfare $W^I(\varphi(p, \beta_I))$ strictly decreases for $\beta_I \geq 0$.*

(iv) *If $[p^*(0) - \bar{p}][p - p^*(0)] > 0$, then social welfare $W^I(\varphi(p, \beta_I))$ strictly increases in $[0, \beta_w(p)]$ and strictly decreases in $(\beta_w(p), \infty)$, where*

$$\beta_w(p) = \frac{R - p - \sqrt{RC/\mu}}{\ln(\mu - \sqrt{C\mu/R})/(\lambda - \mu + \sqrt{C\mu/R})}.$$

This proposition fully characterizes the social welfare as a function of the level of bounded rationality. Figure 5 depicts the different scenarios in Proposition 9 based on the relative magnitude of $p^*(0)$ and $\bar{p}$. By comparing the magnitude of $p^*(0)$ and $\bar{p}$, we discuss two cases. For each case, we depict the range of price p that falls into scenarios (i)–(iv) in Proposition 9. (The case when $p^*(0) = \bar{p}$ is simple and hence not illustrated in the figure.) For the first scenario, the joining fraction at price $p = \bar{p}$ is precisely half, and it is independent of the level of bounded rationality. Thus the social welfare in (i) is not impacted by the level of bounded rationality. For the second scenario, the fraction 0.5 lies *between* the joining fraction induced by the welfare-maximizing price when customers are

Figure 5 **Illustration of the Various Scenarios in Proposition 9**



fully rational and the joining fraction induced by price p when customers' level of bounded rationality is β_I. In this case, increasing the level of bounded rationality makes their "distance" smaller because the equilibrium joining fraction becomes closer to 0.5 as β_I increases, based on Proposition 3, scenarios (ii) and (iii). Thus, the social welfare strictly increases as customers are more boundedly rational. For the third scenario, the joining fraction induced by the welfare-maximizing price when customers are fully rational is either too high or too low compared with the joining probability induced by price p when customers' level of bounded rationality is β_I, so that increasing the level of bounded rationality can only make their "distance" further apart. Therefore, the social welfare strictly decreases in the level of bounded rationality β_I. For the last scenario, the joining probability induced by the welfare-maximizing price when customers are fully rational can be achieved (in the interior). Hence, as the level of bounded rationality increases from zero, the social welfare is "closer" to the optimal social welfare. In this case, the social welfare function is unimodal in the level of bounded rationality, and the first-order condition yields the level of bounded rationality $\beta_w(p)$.

Proposition 9 implies that the social welfare function is unimodal in the level of bounded rationality, as stated in Corollary 2.

Corollary 2. *Social welfare $W^I(\varphi(p,\beta_I))$ is unimodal in the level of bounded rationality β_I for any price p.*

We have demonstrated that the impact of bounded rationality on social welfare depends on the magnitude of the fixed price charged and that the welfare function is unimodal in the level of bounded rationality. The next question we are interested in is, what is the welfare-maximizing price and how does the welfare behave under such a price? We first prove that the social welfare $W^I(\varphi(p,\beta_I))$ is unimodal in price p for any level of bounded rationality β_I (see Lemma EC.4 in the appendix of Huang et al. 2012 for a rigorous justification). Finding the welfare-maximizing price boils down to finding the optimal joining probability φ_w^*. To derive the welfare-maximizing price, we use the first-order condition

$$\partial W^I(\varphi(p,\beta_I))/\partial\varphi(p,\beta_I)=0$$

and obtain

$$\varphi_w^*=\frac{\mu-\sqrt{C\mu/R}}{\lambda},$$

which is the optimal equilibrium joining fraction that induces the optimal social welfare. Suppose this equilibrium point can be achieved in the interior; then it is required that $R\in(C/\mu,\infty)$ if $\mu<\lambda$ and $R\in(C/\mu, C\mu/(\mu-\lambda)^2)$ if $\mu>\lambda$. For cases when the equilibrium point is on the boundary, the problem becomes trivial: if $R\le C/\mu$, then it is socially optimal to keep everybody out of the system; if $R\ge C\mu/(\mu-\lambda)^2$ when $\mu>\lambda$, then it is socially optimal to let everyone join the system.

We are now ready to state the welfare-maximizing price that maximizes the social welfare. To state the result, we first substitute the joining fraction φ_w^* into the equilibrium condition, i.e., Equation (1), and then we obtain the "unconstrained" optimal price by

$$\begin{aligned}p_w^*(\beta_I)&=R-\sqrt{\frac{CR}{\mu}}-\beta_I\ln\frac{\mu-\sqrt{C\mu/R}}{\lambda-\mu+\sqrt{C\mu/R}}\\&=p^*(0)-\beta_I\ln\frac{\varphi_w^*}{1-\varphi_w^*},\end{aligned}$$

which can be negative. The welfare-maximizing price is thus $\max\{0,p_w^*(\beta_I)\}$. Proposition 10 characterizes this welfare-maximizing price and the corresponding social welfare.

Proposition 10. (i) *If $R>4C\mu/(2\mu-\lambda)^2$ when $\beta_I<\beta_w(0)$, where $\beta_w(0)=(R-\sqrt{RC/\mu})/\ln[(\mu-\sqrt{C\mu/R})/(\lambda-\mu+\sqrt{C\mu/R})]$, the price $p=p_w^*(\beta_I)$ is the unique price that maximizes the social welfare, $p_w^*(\beta_I)$ strictly decreases in β_I, and the optimal social welfare is $W^I(p_w^*,\beta_I)=\mu R+C-2\sqrt{\mu RC}$; when $\beta_I\ge\beta_w(0)$, the price $p=0$ is the unique price that yields the maximum social welfare $W^I(0,\beta_I)$.*

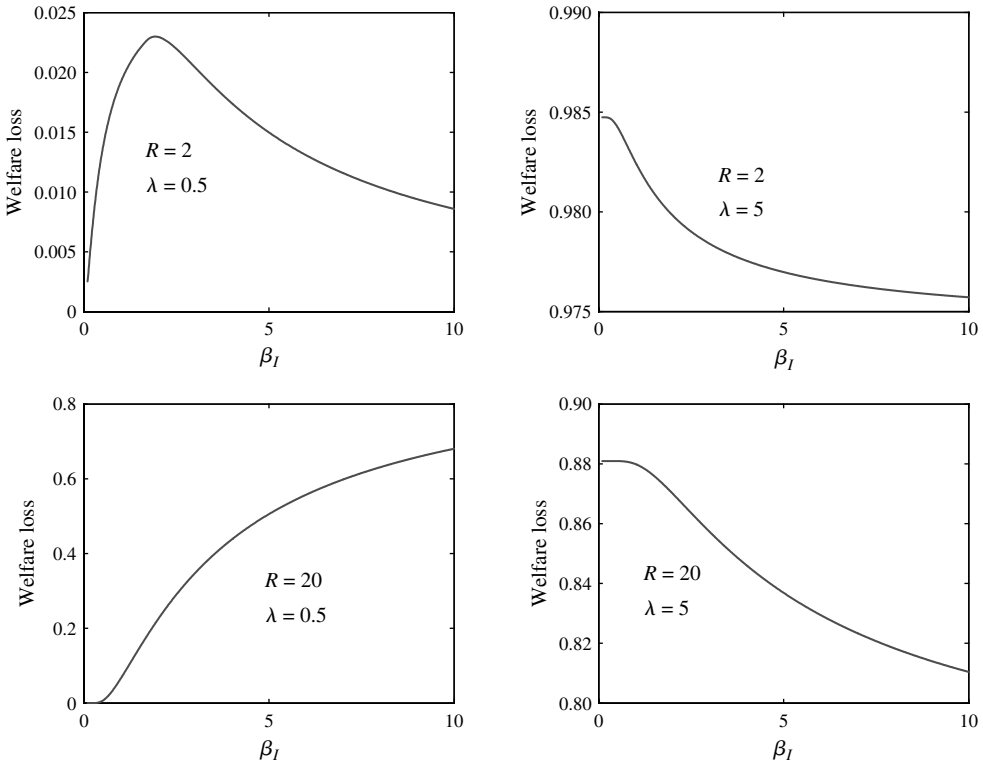
(ii) *If $R\le 4C\mu/(2\mu-\lambda)^2$, the price $p=p_w^*(\beta_I)$ is the unique price that maximizes the social welfare, $p_w^*(\beta_I)$ strictly increases in β_I, and the optimal social welfare is $W^I(p_w^*,\beta_I)=\mu R+C-2\sqrt{\mu RC}$.*

We discuss the implications of Proposition 10 as follows. If $\varphi_w^*=(\mu-\sqrt{C\mu/R})/\lambda>1/2$, then $\ln(\mu-\sqrt{C\mu/R})/(\lambda-\mu+\sqrt{C\mu/R})>0$, which implies that the price $p_w^*(\beta_I)$ is strictly decreasing in level of bounded rationality β_I. This scenario corresponds to Case II depicted in Figure 5, where the equilibrium joining fraction decreases in β_I. In particular, when customers are slightly boundedly rational, the optimal price strictly decreases. The intuition is that the equilibrium joining fraction decreases as the level of bounded rationality increases. To achieve the desired optimal joining fraction φ_w^*, the social planner has to lower the price as the level of bounded rationality increases.

Similarly, if $\varphi_w^*=(\mu-\sqrt{C\mu/R})/\lambda<1/2$, then $\ln(\mu-\sqrt{C\mu/R})/(\lambda-\mu+\sqrt{C\mu/R})<0$, which implies that the price $p_w^*(\beta_I)$ is strictly increasing in level of bounded rationality β_I. This scenario corresponds to Case I depicted in Figure 5, where the equilibrium joining fraction increases in β_I.

The key insight from this proposition is that the first-best social welfare (which is independent of the

Figure 6 Welfare Loss When Hiding the Queue If Bounded Rationality Is Ignored ($C = 1, \mu = 1$)



level of bounded rationality and the arrival rate) can be achieved when either (i) the optimal joining fraction for social welfare maximization is strictly above half and the level of bounded rationality is not too high or (ii) the optimal joining fraction for social welfare maximization is below half.

The intuition for this key insight is the following: in these settings, the social planner can always correct for the boundedly rationality on the part of customers; i.e., the social planner still achieves the same optimal social welfare by charging *appropriate* prices. However, when the optimal joining fraction for social welfare maximization is strictly above half and the level of bounded rationality is sufficiently high, the first-best social welfare cannot be achieved. In other words, when the desired joining fraction is high, to achieve this, the customers have to join with this fraction in equilibrium. However, the customers' joining fraction would be much lower if they are too boundedly rational and too low even if the firm does not charge any price. In this case, higher bounded rationality leads to more social welfare losses. This result stands in contrast to (i) the case of revenue maximization and (ii) the result in Proposition 8 of the visible queue.

Impact of Ignoring Bounded Rationality. Without taking into account customer bounded rationality, the social planner will rationally charge price $p^*(0)$, which is generally different from the welfare-maximizing price $\max\{0, p_w^*(\beta_I)\}$. Similarly, we are interested in the welfare loss

$$\Delta W^I(\beta_I) \equiv (W^I(\max\{0, p_w^*(\beta_I)\}, \beta_I) - W^I(p^*(0), \beta_I))/W^I(p^*(0), \beta_I).$$

For the same example as the revenue maximization case, Figure 6 shows that the welfare loss can be significant (more than 80%, for instance) and is not necessarily monotone with respect to the level of bounded rationality. The nonmonotonicity behavior and its intuitive explanation stems from Proposition 9, where for a fixed price, the social welfare may be nonmonotone in β_I.

5. Discussion

The quantal choice paradigm in the behavioral economics literature posits that people are more likely to select better choices than worse ones but do not necessarily succeed in selecting the very best choice. In this paper, we adopted this framework to model bounded rationality in service systems in the sense that customers lack the capability to perfectly estimate their expected waiting time. We investigated the impact of bounded rationality on the revenue of a profit-maximizing firm, social welfare, and pricing for both invisible and visible queues. From the firm's perspective, higher β_I can lead to lower optimal prices, but it leads to higher optimal prices and higher

revenue when β_I is sufficiently large. With the optimal price, a strictly positive and sufficiently small β results in revenue losses. From the social planner's perspective, there may be strictly positive social welfare losses when β_I is sufficiently large. For visible queues with a fixed price, we prove that a strictly positive and sufficiently small β can lead to strict social welfare improvement, and we provide a simple inequality under which this improvement happens. With the optimal prices, however, bounded rationality decreases social welfare. We demonstrate that ignoring bounded rationality may result in significant revenue and social welfare loss. Our study contributes to the behavioral operations management literature by demonstrating the impact of behavioral factors in service settings.

5.1. Experimental Design

We hope our theoretical study helps develop a desirable theoretical–empirical feedback loop in behavioral operations management. To capture bounded rationality on the customers' part, in our model the customers have imperfect estimates of the expected waiting time. Loosely speaking, the standard deviation of the estimate of expected waiting time is used as a proxy for the level of bounded rationality.

The first step for such a study will be to estimate the level of bounded rationality β for visible and invisible queues in carefully controlled experiments. This could be done by assigning the experimental parameters to subjects and then observing their joining fractions. Standard maximum likelihood estimation would yield the estimate for β, and we can test whether it is statistically significantly different from zero. As we mentioned in §2, McKelvey and Palfrey (1995) actually follow this approach in bimatrix game settings and Bajari and Hortacsu (2005) in auction settings. Kremer and Debo (2012) have already carried out experimental studies in service systems and found that the model of bounded rationality fits the experimental data very well.

There are a variety of interesting conjectures of customer behavior that could be tested or explored along the line of Kremer and Debo (2012). It is of interest to estimate β for different queue configurations. For example, Larson (1987) and Maister (1985) discuss the impact of the queue environment on the customer waiting time perception. Specifically, they conjecture that eliminating empty time significantly reduces customer perception of the length of waiting time and that explained waits are shorter than unexplained waits. We conjecture that customers may have different levels of bounded rationality for these different queue environments. Some conjectures worth exploring are the following: Are people prone to estimation mistakes/errors (and hence bounded rationality) depending on the queue structure, for instance, with one long queue versus many short queues? Are more educated (or more knowledgeable) people less boundedly rational? Answering these questions not only deepens our understanding of customer behavior but also helps managers better operate service systems.

It is interesting yet challenging to empirically quantify the loss that the firm incurs when it disregards bounded rationality. Although the arrival rate λ, service rate μ, and price p are easy to estimate, the reward R, level of bounded rationality β, and cost C are a little more difficult to identify, given the possible heterogeneity and interaction among these parameters. These challenges can be overcome using instrumental analysis or other structural estimation methods. Once these parameters are estimated, we can compute the loss that the firm incurs.

Finally, recall that we have theoretically proven that bounded rationality can improve social welfare. Searching for empirical evidence that in systems where bounded rationality exists indeed improves the social welfare would be highly valuable.

5.2. Managerial Insights

There are several implications for how service systems should be managed. First, the study demonstrates the importance of accounting for bounded rationality in pricing the service. In particular, as β increases above a certain threshold, the revenue and social welfare loss of not accounting for it is large. Second, there are settings where the system manager can reduce the ambiguity (or difficulty) associated with the process of estimating the waiting time and potentially reduce the variance in the estimation. It is interesting to note that because customers are self-interested, when it comes to welfare maximization (e.g., public systems such as the Department of Motor Vehicles and traffic on a road network), reducing β to zero might not be the right step because bounded rationality can actually improve social welfare. However, when the service provider can optimize the price as well as reduce the level of bounded rationality significantly, then the system performance (both in terms of revenue and social welfare) can be dramatically improved.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/msom.1120.0417.

Acknowledgments

The authors thank the three anonymous referees, the associate editor, and editor Stephen Graves for many helpful comments and suggestions that improved the paper enormously.

References

Afèche P (2004) Incentive-compatible revenue management in queueing systems: Optimal strategic delay and other delay tactics. Working paper, University of Toronto, Toronto.

Anderson SP, de Palma A, Thisse J-F (1992) *Discrete Choice Theory of Product Differentiation* (MIT Press, Cambridge, MA).

Ariely D (2009) The end of rational economics. *Harvard Bus. Rev.* 87(7/8):78–84.

Arkes HR, Hammond KR, eds. (1985) *Judgment and Decision Making: An Interdisciplinary Reader* (Cambridge University Press, Cambridge, UK).

Bajari P, Hortacsu A (2001) Auction models when bidders make small mistakes: Consequences for theory and estimation. Working paper, Stanford University, Stanford, CA.

Bajari P, Hortacsu A (2005) Are structural estimates of auction models reasonable? Evidence from experimental data. *J. Political Econom.* 113(4):703–741.

Basov S (2009) Monopolistic screening with boundedly rational consumers. *Econom. Record* 85(S1):S29–S34.

Bearden JN, Rapoport A, Seale DA (2005) Entry times in queues with endogenous arrivals: Dynamics of play on the individual and aggregate levels. Rapoport A, Zwick R, eds. *Experimental Business Research*, Vol. II (Springer, Berlin), 201–221.

Bendoly E, Donohue K, Schultz K (2006) Behavioral operations management: Assessing recent findings and revisiting old assumptions. *J. Oper. Management* 24(6):737–752.

Bendoly E, Croson R, Goncalves P, Schultz K (2009) Bodies of knowledge for research in behavioral operations. *Production Oper. Management* 19(4):434–452.

Cason TN, Reynolds SS (2005) Bounded rationality in laboratory bargaining with asymmetric information. *Econom. Theory* 25(3):553–574.

Chen HC, Friedman JW, Thisse JF (1997) Boundedly rational Nash equilibrium: A probabilistic choice approach. *Games Econom. Behav.* 18(1):32–54.

Conlisk J (1996) Why bounded rationality? *J. Econom. Literature* 34(2):669–700.

Davis AM (2011) An experimental investigation of pull contracts. Working paper, Pennsylvania State University, University Park.

Geigerenzer G, Selten R (2001) *Bounded Rationality: The Adaptive Toolbox* (MIT Press, Cambridge, MA).

Gino F, Pisano G (2008) Toward a theory of behavioral operations. *Manufacturing Service Oper. Management* 10(4):676–691.

Goeree JK, Holt CA, Palfrey TR (2002) Quantal response equilibrium and overbidding in private-value auctions. *J. Econom. Theory* 104(1):247–272.

Hassin R (1986) Consumer information in markets with random products quality: The case of queues and balking. *Econometrica* 54(5):1185–1195.

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems* (Kluwer Academic Publishers, Norwell, MA).

Hey J, Orme C (1994) Investigating generalizations of expected utility theory using experimental data. *Econometrica* 62(6): 1291–1326.

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

Hogarth R (1980) *Judgment and Choice: Psychology of Decision* (John Wiley & Sons, New York).

Hsu VN, Xu SH, Jukic B (2009) Optimal scheduling and incentive compatible pricing for a service system with quality of service guarantees. *Manufacturing Service Oper. Management* 11(3):375–396.

Huang T, Allon G, Bassamboo A (2012) Technical report to "Bounded rationality in service systems." http://www.kellogg.northwestern.edu/research/operations/workingpapers.htm.

Kahneman D, Slovic P, Tversky A, eds. (1981) *Judgment Under Uncertainty: Heuristic and Biases* (Cambridge University Press, Cambridge, UK).

Knudsen NC (1972) Individual and social optimization in a multiserver queue with a general cost-benefit structure. *Econometrica* 40(3):515–528.

Kremer M, Debo L (2012) Herding in a queue: A laboratory experiment. Chicago Booth Research Paper 12-28, Chicago Booth School of Business, Chicago.

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

Larson RC (1987) Perspectives on queues: Social justice and the psychology of queueing. *Oper. Res.* 35(6):895–905.

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

Lippman SA, Stidham S Jr (1977) Individual versus social optimization in exponential congestion systems. *Oper. Res.* 25(2): 233–247.

Loomes G, Moffatt P, Sudgen R (2002) A microeconometric test of alternative stochastic theories of risky choice. *J. Risk Uncertainty* 24(2):103–130.

Luce RD (1959) *Individual Choice Behavior: A Theoretical Analysis* (John Wiley & Sons, New York).

Maister D (1985) The psychology of waiting lines. Czepiel JA, Solomon MR, Suprenant C, eds. *The Service Encounter*, Chap. 8 (Lexington Books, New York), 113–126.

Mallard G (2011) Modelling cognitively bounded rationality: An evaluative taxonomy. *J. Econom. Surveys* 26(4):674–704.

Mattsson L-G, Weibull JW (2002) Probabilistic choice and procedurally bounded rationality. *Games Econom. Behav.* 41(1):61–78.

McFadden D (1974) Conditional logit analysis of qualitative choice behavior. Zarembka P, ed. *Frontiers in Econometrics* (Academic Press, New York), 105–142.

McKelvey RD, Palfrey TR (1995) Quantal response equilibria for normal form games. *Games Econom. Behav.* 10(1):6–38.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.

Nisbett R, Ross L (1980) *Human Inference: Strategies and Shortcomings in the Social Judgment* (Prentice-Hall, Englewood Cliffs, NJ).

Payne JW, Bettman JR, Johnson EJ (1992) Behavioral decision research: A constructive processing perspective. *Annual Rev. Psych.* 43:87–131.

Pitz G, Sachs NJ (1984) Judgment and decision: Theory and application. *Annual Rev. Psych.* 35:139–162.

Plambeck EL, Wang Q (2010) Implications of hyperbolic discounting for optimal pricing and information management in service systems. Working paper, Stanford University, Stanford, CA.

Rapoport A, Stein WE, Parco JE, Seale DA (2004) Equilibrium play in single-server queues with endogenously determined arrival times. *J. Econom. Behav. Organ.* 55(1):67–91.

Seale DA, Parco JE, Stein WE, Rapoport A (2005) Joining a queue or staying out: Effects of information structure and service time on arrival and staying out decisions. *Experiment. Econom.* 8(2):117–144.

Simon HA (1955) A behavioral model of rational choice. *Quart. J. Econom.* 69(1):99–118.

Simon HA (1957) *Models of Man* (John Wiley & Sons, New York).

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

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

Talluri KT, van Ryzin GJ (2004) *The Theory and Practice of Revenue Management* (Springer-Verlag/Kluwer Academic Publishers, New York).

Thurstone LL (1927) A law of comparative judgment. *Psych. Rev.* 34(4):273–286.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Van Mieghem JA (2000) Price and service discrimination in queuing systems: Incentive compatibility of $Gc\mu$ scheduling. *Management Sci.* 46(9):1249–1267.

Waksberg A, Smith A, Burd M (2009) Can irrational behaviour maximise fitness? *Behav. Ecol. Sociobiol.* 63(3):461–471.

Yechiali U (1971) On optimal balking rules and toll charges in a $GI/M/1$ queuing process. *Oper. Res.* 19(2):349–370.

Yechiali U (1972) Customers' optimal joining rules for the $GI/M/s$ queue. *Management Sci.* 18(7):434–443.