Priority Queues and Consumer Surplus

Martin A. Lariviere
Northwestern University

September 4, 2018

Abstract

We examine whether priority queues benefit or hurt customers in a setting in which customers are privately informed of their per-unit-time waiting cost. Implementing a priority queue thus means posting a menu of expected waits and out-of-pocket prices that are incentive compatible. Whether priorities increase or decrease consumer surplus relative to first-in, first-out service depends on the model of customer utility and on the distribution of customer waiting costs. If all customers have the same value of the service independent of their waiting costs, priorities essentially always lower consumer surplus. If a customer's value of the service is an increasing function of her waiting cost, priorities lower surplus if the distribution of waiting costs has a decreasing mean residual life. If the mean residual life is increasing, then priorities make consumers better off. We show that the results across utility models are linked by an elasticity measure. If an appropriate measure of waiting cost is elastic, consumer surplus falls with priorities. We also explore how priorities impact individual customers and show that they potentially make all customers worse off. It is possible that low priority customers may pay a higher out-of-pocket price than they would under first-in, first-out service.

1. Introduction

Are consumers better off when a queue is run on a first-come, first-served basis or when a priority scheme is used? The answer is not immediately clear. On the one hand, customers in many instances object to deviations from first-in, first-out (FIFO) service. In 2014, British telecom provider EE introduced a system in

which customers could receive priority service for a seemingly nominal fee (50p) when contacting the firm's call center. Customers responded by threatening to switch carriers (Johnston, 2014). When Belgian amusement park Walibi launched an expensive ticket that allowed holders to jump the queue at all rides, even government ministers chimed in opposing the program (Flandersnews.be 2013). A V.I.P. ticket program at Universal Studios Hollywood that allowed holders to jump to the front of the line (along with other perks) also garnered negative reactions (Barnes, 2013). In 2010, the U.S. Patent and Trademark Office proposed offering expedited review of patent filings for an extra fee, a move which some argued would harm independent inventors (Schatz, 2010). Finally, many proponents of net neutrality fear that absent government intervention service providers will implement "paid prioritization" (Puzzanghera, 2017).

On the other hand, classic results in queuing theory show that priorities are an essential tool in managing systems in which jobs incur delay costs (e.g., Cox and Smith, 1961). As such, one might well expect priorities to benefit customers. In the words of Larson (2012)

> "[A priority queue] is an example of 'demand management', in which price of a service is matched to demand. ... it is simply another manifestation of matching supply and demand by market pricing. Often those paying the highest price subsidise others who can then enjoy the service, though perhaps with a bit of temporal inconvenience."

We examine the tension between customer apprehension of priority queues and manager affinity for them from the perspective of consumer surplus. That is, we ask whether priorities make consumers as a collective better off.

There are obviously settings where priorities benefit consumers. An emergency department, for example, has patients for whom each additional minute of delay raises the probability of a tragic outcome. Further, these patients can be separated from those with less critical needs through a triage process. Arrivals are not opting for a particular class of service based on posted prices and expected waits. We do not consider such markets. We are interested in settings like as a customer service call center or an amusement park where the class of service to seek is a choice of the customer and depends on the price charged as well as the expected wait. In our model, customers draw a type from a continuous distribution. A customer's type corresponds to her per-unit-time cost of waiting and is privately observed. We consequently focus on systems managed via incentive compatible prices. One can envision the queue as being governed by a menu of posted prices and waits.

A customer given her type then chooses whether to patronize the service system and if so which class of service to buy (assuming priorities are used).

We consider two models of customer utility. In the fixed values model, all customers have the same value of service (gross of any delay cost) that is independent of their waiting costs. In the increasing values model, a customer's value of the service is an increasing function of her type. Consequently, in the increasing values case, there exists a cutoff such that all types above that cutoff patronize the service. In the fixed values case, there is similarly a cutoff type but now it is an upper bound and all customers below the cutoff seek service. Both of these models have precedents in the literature. Fixed values are used in Ghanem (1975); Gilland and Warsing (2009); and Gavirneni and Kulkarni (2013). Variations on our increasing values model are used by Afèche and Mendelson (2004); Doroudi et al. (2013); Nazerzadeh and Randhawa (2017); Gurvich et al. (2018).[1]

For both of these settings, we evaluate the impact of priorities by holding the total number of customers served constant and asking whether consumer surplus is higher under FIFO service or a simple static priority scheme with two classes (which we refer to as HiLo priorities). We do not explicitly consider optimization. That is, we do not ask what priority scheme would maximize consumer surplus. Rather, we seek conditions such that the move from FIFO to any HiLo priority scheme would raise or lower consumer surplus. We show the following:

1. **Priorities increase revenue and social welfare.** For either model of customer utility, priorities dominate FIFO in the sense that an arbitrarily designed HiLo scheme results in higher revenue and social welfare than a FIFO queue with the same arrival rate.

2. **Gains in revenue and welfare may come at the expense of customers.** Under the fixed values model, HiLo essentially always lowers consumer surplus. Under the increasing values model, consumers may be helped or harmed by the imposition of priorities. If the distribution of customer types has a decreasing mean residual life function, surplus falls. If the mean residual life is increasing, surplus is increasing. With slightly more restrictive assumptions on the type distribution, we can extend these results to priority schemes with an arbitrary number of classes.

[1] In these papers, the customer's type is a value for the service and the waiting cost is then a function of the service value. We have reversed this relationship to allow for an easier comparison with the fixed values model.

3. **When consumer surplus is very sensitive to throughput, surplus falls.** We are able to link the fixed values and increasing values results through an elasticity measure. Under either setting, if delay costs are elastic with respect to the throughput of the queue, surplus falls.

4. **Variability matters.** How much consumers win or lose with a change in the service discipline depends in part on the distribution of customer types. In the fixed values model, customer are better off (in the sense of incurring a relatively smaller loss when moving from FIFO to HiLo) if the type distribution is less variable. These results are reversed in the increasing values model.

5. **Priorities create winners and losers – or just losers.** In the increasing values model, those with very high waiting costs benefit from the move to priorities while patient customers are worse off. In the fixed values model, however, it is possible for all customers to be worse off under priorities. In particular, it is possible that the price for low class service is higher than the price for the FIFO queue. Instead of having their costs subsidized by high priority patrons, these customers are both paying more and receiving worse service.

In what follows, we first review the relevant literature and then cover the basics of the model. Section 4 examines how priorities affect social welfare and revenue. Section 5 presents our key results on when priorities benefit or hurt customers. Section 6 covers several extensions. Proofs are in the Appendix.

2. Literature review

The concept of consumer surplus has a long history in the field of economics (Dupuit, 1844). One line of work in this area looks at how third-degree price discrimination impacts customers. Third-degree price discrimination assumes that a firm can offer the same product at different prices to distinct segments (e.g., students versus non-students). Cowan (2012) presents conditions under which consumers benefit if the seller switches from uniform pricing (i.e., using the same price in all markets) to differentiated pricing when the segments are differentiated by their demand elasticities. Chen and Schwartz (2015) perform a similar analysis assuming that the seller has different costs to serve each segment. This work

differs from ours on two dimensions. First, the work on third-degree price discrimination assumes that segments are exogenously specified and membership can be verified (i.e., a customer does nor does not have a student ID). In our model, customers decide for themselves which class of service to buy so prices must be set to be incentive compatible. Second, our seller is offering both different prices and different levels of service.

Recent work in the Operations Management literature has also examined consumer surplus. Stamatopoulos et al. (2017) consider a retailer managing inventory with an economic order quantity cost structure and demand that varies with the retail price. They show that allowing the retailer to dynamically adjust the price with its available inventory increases both retailer profits and consumer surplus. Chen and Gallego (2018) examine the impact of dynamic pricing on social welfare and consumer surplus in standard revenue management models and present conditions under which revenue maximizing policies benefit consumers. They also briefly discuss a queuing model motivated by transportation settings. However, this model does not explicitly model customer waiting costs. These papers differ from our work since they focus on dynamically adjusting pricing based on the state of the system and consider only one offering. We consider settings in which prices and expected waits for possibly multiple classes of service are set independently of current congestion.

Comparisons of revenue maximization and social optimality in queues have been studied since at least Naor (1969). However, relatively little attention has been given to consumer surplus. Indeed, the only mentions of consumer surplus in a standard survey on the economics of queue (Hassin and Haviv, 2003) are identifying settings in which a revenue maximizing firm is able to expropriate all consumer surplus so revenue maximization results in welfare maximization.

An exception is Gurvich et al. (2018). Using a linear version of the increasing values framework, they consider a decision maker choosing how much of the market to cover, how coarse a priority scheme to offer, and how to classify customers into priority classes. They then compare the choices of a revenue maximizer with those of the social planner. In a limiting regime, the revenue maximizer and the social planner offer essentially the same levels of coverage and both can offer just two priority classes. However, they classify customer differently and that this classification depends on the mean residual life function of the type distribution, which represents consumer surplus per customer. They also show that priorities can increase or decrease consumer surplus relative to FIFO depending on the failure rate of the type distribution.

We extend their results in several ways. Working with two priority classes, we are able to consider a more general relationship between service value and waiting cost. We have a simpler approach to proving our results and employ a somewhat less restrictive condition. We have new results on how the change in surplus from implementing priorities depends on characteristics of the type distribution. Finally, we also examine the fixed value model.

3. Model basics

Here we first layout our model of the service system and then present our two models of consumers.

3.1. The service setting

We consider a service modeled as a Markovian queue. Customers arrive according to a Poisson process with mean Λ. Services times are independent draws from an exponential distribution with mean one. We will generally be vague about the number of servers; the bulk of our results cover both single and multi-server queues. However, we will assume that Λ exceeds available capacity so customers must be turned away. All numerical results will assume a single server.

The queue is managed under one of two service disciplines, either FIFO or high-low priority (HiLo). We use the latter as shorthand for a static, two-class priority scheme. Let W_F denote the expected wait under FIFO while W_H and W_L respectively denote the expected waits for high-priority and low-priority customers under HiLo. Similarly, define the arrival rates as λ_F for FIFO and λ_H and λ_L for HiLo. Our primary interest will be in how consumer surplus varies as we move from FIFO to HiLo assuming coverage is unchanged, i.e., $\lambda_F = \lambda_H + \lambda_L$.

The current state of the system is not visible to arriving customers. Customers must then choose whether to purchase service (and what class of service if priorities are used) based on a posted menu of prices and expected delays. We assume that under either service discipline that the system offers the minimum feasible waits given the arrival rates. Thus the system employs a work-conserving policy and does not insert any strategic delay (Afèche, 2013). We consequently have

$$W_F = \frac{\lambda_L}{\lambda_F} W_L + \frac{\lambda_H}{\lambda_F} W_{H.} \tag{1}$$

Note that (1) holds whether preemptive or non-preemptive priorities are used under HiLo. All numerical results will assume preemptive priorities.

3.2. Consumers

We consider two models of consumer utility. Under both, arriving customers independently draw a type t from a known, continuous distribution. A customer's realized type is her per-unit-time cost of waiting. It is privately observed; whoever is managing the queue consequently cannot simply route customers based on their types. Rather, they must post a menu of incentive compatible prices and waits to induce customers to make the appropriate choices.

The models differ in how a customer's valuation of the service relates to her type. In the *Fixed Values* model, all consumers have the same valuation for the service. In the *Increasing Values* model, a customer's valuation is an increasing function of her type. We discuss each in turn.

3.2.1. Fixed values

In the fixed values model, customers draw their types from distribution $F(t)$ with density $f(t)$. Let $\bar{F}(t) = 1 - F(t)$. We assume that $F(t)$ is continuous on $(0, \Omega)$ with a finite mean. All customers value the service (gross of waiting costs) at V. A customer with realized type t who expects a wait of W and pays an out of pocket price of p then has an expected utility of

$$U(p, W|t) = V - p - tW.$$

For a fixed p and W, utility is decreasing in the customer's type. Consequently, if a type t customer buys, it must be the case that any customer with type $t' < t$ must also buy. Further, it is straightforward to show that if customers with types t_1 and t_2 such that $t_1 < t_2$ purchase the same class of service under HiLo, any customer with type $t_3 \in (t_1, t_2)$ must also buy the same class. Finally, if customers at t_1 and $t_2 > t_1$ buy different classes of service under HiLo, it must be the case that t_1 buys low priority service at a low price while t_2 buys high priority service at a high price.

To determine incentive compatible prices for some targeted arrival rates, define $t_L = F^{-1}(\lambda_L/\Lambda)$ and $t_H = F^{-1}((\lambda_L + \lambda_H)/\Lambda)$. We then have $\lambda_F = \Lambda F(t_H)$, $\lambda_H = \Lambda(F(t_H) - F(t_L))$, and $\lambda_L = \Lambda F(t_L)$. Under either FIFO or HiLo, a customer with realized type t_H is indifferent to buying. The price for FIFO service must be $p_F = V - t_H W_F$, and the price for high-priority service under HiLo must be $p_H = V - t_H W_H$. A customer with realized type t_L is indifferent between the two types of service, so $p_L = p_H - t_L(W_L - W_H)$. Note that here and elsewhere

we suppress the dependence of the expected waits on the arrival rates and hence on t_L and t_H for notational convenience.

3.2.2. Increasing values

Let $G(t)$ denote the type distribution for the increasing values model with $g(t)$ and the $\bar{G}(t)$ as the corresponding density and survival function. $G(t)$ is continuous on $(0, \Omega)$ with a finite mean. In the increasing values model, a customer's valuation of the service $\hat{V}(t)$ is an increasing function of the customer's type such that $\hat{V}(t) > 0$ for $t > 0$. (For clarity we will use "hats" on notation associated with the increasing values model.) Additionally, we assume that $\hat{V}(t)/t$ is increasing in t. A type t customer's expected utility given an out of pocket price p and expected wait W is then

$$\hat{U}(p, W|t) = \hat{V}(t) - p - tW.$$

Given our assumptions, we have that utility is *increasing* in t if a type-t customer enjoys a non-negative utility from puchasing. Consequently, if a customer with realized type t buys so must a customer at $t' > t$. Again, it is straightforward to show that customers buying a particular class of service under HiLo must form an interval of types and that those with higher waiting costs buy high-priority service.

To determine prices, define $\hat{t}_L = \bar{G}^{-1}\left(\left(\hat{\lambda}_L + \hat{\lambda}_H\right)/\Lambda\right)$ and $\hat{t}_H = \bar{G}^{-1}\left(\hat{\lambda}_H/\Lambda\right)$. We then have $\hat{\lambda}_H = \Lambda\bar{G}\left(\hat{t}_H\right)$ and $\hat{\lambda}_L = \Lambda\left(\bar{G}\left(\hat{t}_L\right) - \bar{G}\left(\hat{t}_H\right)\right)$ under HiLo and $\hat{\lambda}_F = \Lambda\bar{G}\left(\hat{t}_L\right)$ under FIFO. A customer with realized type $\hat{t}_L$ is indifferent to buying under both regimes. The price for FIFO service must be $\hat{p}_F = \hat{V}\left(\hat{t}_L\right) - \hat{t}_L W_F$ while the price for low-priority service must $\hat{p}_L = \hat{V}\left(\hat{t}_L\right) - \hat{t}_L W_L$. A type $\hat{t}_H$ customer is indifferent between the two classes of service, yielding $\hat{p}_H = \hat{p}_L + \hat{t}_H\left(W_L - W_H\right)$.

3.2.3. Comparing the models

As noted above, both the fixed values model and the increasing values model have substantial precedents in the literature. Indeed, they are arguably the two standard consumer models used for studying revenue management in queueing settings.

The models result in different problem structures. With fixed values, customers with higher waiting costs create less value for the system, and it is the customer with the highest waiting cost in a particular class that dictates the price for that grade of service. This is reversed in the increasing values models. High waiting

costs move in lock step with high valuations; it is now those with high types that create the most value while those with low waiting costs end up determining prices.

The models also represent fundamentally different settings. A more general formulation would endow customers with two dimensional types so that both waiting costs and service valuations are drawn at random. One could then allow some degree of correlation between waiting costs and valuations. Independence and positive correlation would be obvious cases of interest. Our two models of consumers capture simplified versions of these settings. Having a fixed service value for all waiting costs is the simplest model of independence while having values increasing deterministically with waiting costs is the simplest model of positive correlation.

4. Social welfare and revenue

Before turning to consumer surplus, we consider how social welfare and revenue change as one moves from FIFO to HiLo. For the fixed values model, let $\rho_{HL}(t_L, t_H)$ denote revenue under HiLo priorities when all customer with types below t_H are admitted and those with types below t_L buy low-priority service. Similarly, let $\omega_{HL}(t_L, t_H)$ denote social welfare under HiLo.

$$\begin{aligned}
\rho_{HL}(t_L, t_H) &= \lambda_L p_L + \lambda_H p_H \\
&= \Lambda\left[F(t_H) V - F(t_H) t_H W_H - F(t_L) t_L (W_L - W_H)\right] \qquad (2) \\
\omega_{HL}(t_L, t_H) &= \lambda_L \frac{\int_0^{t_L} (V - tW_L) f(t)\, dt}{F(t_L)} + \lambda_H \frac{\int_{t_L}^{t_H} (V - tW_H) f(t)\, dt}{F(t_H) - F(t_L)} \\
&= \Lambda\left[F(t_H) V - W_H \int_0^{t_H} t f(t)\, dt - (W_L - W_H) \int_0^{t_L} t f(t)\, dt\right] \qquad (3)
\end{aligned}$$

Let $\rho_{FF}(t_H)$ and $\omega_{FF}(t_H)$ respectively denote the complementary values for revenue and social welfare under FIFO. These are special cases of HiLo with $t_L = 0$ so that W_H would equal W_F.

We similarly define $\hat{\rho}_{HL}\left(\hat{t}_L, \hat{t}_H\right)$ and $\hat{\omega}_{HL}\left(\hat{t}_L, \hat{t}_H\right)$ as revenue and social welfare

under HiLo priorities in the increasing values model.

$$\begin{aligned}
\hat{\rho}_{HL}\left(\hat{t}_L,\hat{t}_H\right) &= \hat{\lambda}_L\hat{p}_L+\hat{\lambda}_H\hat{p}_H \\
&= \Lambda\left[\bar{G}\left(\hat{t}_L\right)\hat{V}\left(\hat{t}_L\right)-\bar{G}\left(\hat{t}_L\right)\hat{t}_L W_L+\bar{G}\left(\hat{t}_H\right)\hat{t}_H\left(W_L-W_H\right)\right] \quad (4) \\
\hat{\omega}_{HL}\left(\hat{t}_L,\hat{t}_H\right) &= \hat{\lambda}_H\frac{\int_{\hat{t}_H}^{\Omega}\left(\hat{V}(t)-tW_H\right)g(t)\,dt}{\bar{G}\left(\hat{t}_H\right)}+\hat{\lambda}_L\frac{\int_{\hat{t}_L}^{\hat{t}_H}\left(V-tW_L\right)g(t)\,dt}{\bar{G}\left(\hat{t}_L\right)-\bar{G}\left(\hat{t}_H\right)} \\
&= \Lambda\left[\tilde{V}(t)-W_L\textstyle\int_{\hat{t}_L}^{\Omega}tg(t)\,dt+\left(W_L-W_H\right)\int_{\hat{t}_H}^{\Omega}tg(t)\,dt\right], \quad (5)
\end{aligned}$$

where $\tilde{V}(t)=\int_{\hat{t}_L}^{\Omega}\hat{V}(t)\,g(t)\,dt$. Define $\hat{\rho}_{FF}\left(\hat{t}_L\right)$ and $\hat{\omega}_{FF}\left(\hat{t}_L\right)$ as the complementary FIFO values found by setting $\hat{t}_H$ to Ω so that W_L would equal W_F.

We say a HiLo priority scheme is nondegenerate if it has a positive arrival rates to both classes. This implies $t_H>t_L>0$ in the fixed values case and $\Omega>\hat{t}_H>\hat{t}_L$ in the increasing values case.

Proposition 1. *In both the fixed values model and the increasing values model, revenue and social welfare are higher under any nondegenerate HiLo priority scheme than under the corresponding FIFO service discipline with the same total arrival rate, i.e.,* $\rho_{HL}(t_L,t_H)>\rho_{FF}(t_H)$, $\omega_{HL}(t_L,t_H)>\omega_{FF}(t_H)$, $\hat{\rho}_{HL}\left(\hat{t}_L,\hat{t}_H\right)>\hat{\rho}_{FF}\left(\hat{t}_L\right)$, *and* $\hat{\omega}_{HL}\left(\hat{t}_L\hat{t}_H\right)>\hat{\omega}_{FF}\left(\hat{t}_L\right)$.

This result does not depend on any sense of optimization. It does not, for example, assert that the revenue maximizing priority scheme is preferable to FIFO. Rather, it shows that *any* nondegenerate HiLo scheme trumps FIFO with respect to both revenue and social welfare. Intuitively, priorities allow for a more efficient allocation of waiting costs. With respect to social welfare, the gross value created by the system is the same under HiLo and FIFO since the total arrival rate is fixed. Letting those with high waiting cost jump the queue reduces the total waiting cost the system incurs. The revenue the system generates also increases with a move to priorities. Priorities allow charging those at the front of the line more than under FIFO and that higher revenue always more than offsets any discounts given to those at the back of the line.

Before moving on to consider consumer surplus, we note that both revenue and social welfare under the fixed value model as well as social welfare under the increasing values model are decreasing in all expected waits. Our assumption

then that the system always offers the minimum feasible wait is therefore fairly innocuous; any system that imposed additional delay would do better with respect to welfare and revenue by eliminating the unnecessary wait. The same will be true for revenue under the increasing values model if one allows a few technical assumptions. See, for example, Gurvich et al. (2018).

5. How priorities affect consumers

We now turn to how moving from FIFO to HiLo impacts customers. For the fixed values model, let $\phi_{HL}(t_L, t_H)$ denote consumer surplus under HiLo priorities when all customer with types below t_H are admitted and those with types below t_L buy low-priority service while $\phi_{FF}(t_H)$ represents surplus under FIFO if all types below t_H patronize the system. We have

$$\begin{aligned}\phi_{HL}(t_L, t_H) &= \omega_{HL}(t_L, t_H) - \rho_{HL}(t_L, t_H) \\ &= \Lambda\left[W_L\eta(t_L) + W_H(\eta(t_H) - \eta(t_L))\right] \\ \phi_{FF}(t_H) &= \omega_{FF}(t_H) - \rho_{FF}(t_H) = \Lambda W_F\eta(t_H),\end{aligned}$$

where $\eta(t) = tF(t) - \int_0^t xf(x)\,dx$. The corresponding functions for the increasing values model are

$$\begin{aligned}\hat{\phi}_{HL}(\hat{t}_L, \hat{t}_H) &= \hat{\omega}_{HL}(\hat{t}_L, \hat{t}_H) - \hat{\rho}(\hat{t}_L, \hat{t}_H) \\ &= \Lambda\left[\bar{V}(\hat{t}_L) - \left(W_H\hat{\eta}(\hat{t}_H) + W_L(\hat{\eta}(\hat{t}_L) - \hat{\eta}(\hat{t}_H))\right)\right] \\ \hat{\phi}_{FF}(\hat{t}_L) &= \hat{\omega}_{FF}(\hat{t}_L) - \hat{\rho}_{FF}(\hat{t}_L) = \Lambda\left[\bar{V}(\hat{t}_L) - W_F\hat{\eta}(\hat{t}_L)\right],\end{aligned}$$

where $\hat{\eta}(t) = \int_t^\Omega xg(x)\,dx - t\bar{G}(t)$ and $\bar{V}(t) = \tilde{V}(t) - \hat{V}(t)\bar{G}(t)$.

There are both contrasts and similarities in the drivers of consumer surplus between the two models. A notable difference is that consumer welfare in the fixed values case does not depend on V, i.e., on how much consumers value service. Intuitively, if customers were not delay sensitive, a seller would be able to capture all social welfare by pricing at V. Once consumers incur delay costs, prices are set as a discount off of their service valuation, i.e., consumer surplus is only positive because the seller must compensate them for their wait and cannot perfectly discriminate among them. Consequently, surplus for a given service discipline is increasing in the expected wait.

Under increasing values, how customers value service matters and consumers are better off if they collectively have higher valuations for any type realization

(e.g., if $\hat{V}(t) = \theta\tau(t)$ for $\theta > 0$ and $\tau(t)$ positive and increasing, then surplus would be increasing in θ). Consumer surplus is now a discount off $\bar{V}(t)$ and thus is decreasing in expected waits for a given service discipline.

While expected waits have different directional implications in the two models, there is a similarity in the information conveyed by the weighting parameters $\eta(t)$ and $\hat{\eta}(t)$. Too see this, define $\mu(t) = \eta(t)/F(t)$ and $\hat{\mu}(t) = \hat{\eta}(t)/\bar{G}(t)$, which are, respectively, the reverse mean residual life (RMRL) of $F(t)$ and the mean residual life (MRL) of $G(t)$. In the reliability literature, MRL is the expected life beyond time t of a component that has lasted until time t. The RMRL, in contrast, is the expected time that a component has been inactive given that its failure is discovered at time t.[2] While their interpretation in a reliability context differ, here the two functions are playing essentially identical rolls. In the fixed values model, customers join if their type is below some level so the RMRL represents the gap in per-unit-time waiting costs between the marginal agent that defines the price for a class of service and the average customer who joins that class. In the increasing values model, customers join a class when their type is sufficiently high implying that the MRL is also the average gap in per-unit-time waiting costs.

Note that $\eta(t)$ is always increasing in t while $\hat{\eta}(t)$ is decreasing in t. Both, however, are increasing if we work with the throughput of customers, i.e., if we express the delay weighting factors as a function of λ instead of cutoff values. Let $\kappa(\lambda) = \eta(F^{-1}(\lambda/\Lambda))$ and $\hat{\kappa}(\lambda) = \eta(G^{-1}(1-\lambda/\Lambda))$ and the let $\varepsilon(\lambda)$ and $\hat{\varepsilon}(\lambda)$ be the corresponding elasticities of $\kappa(\lambda)$ and $\hat{\kappa}(\lambda)$, i.e.,

$$\varepsilon(\lambda) = \frac{\lambda\kappa'(\lambda)}{\kappa(\lambda)} = \frac{1}{\mu(F^{-1}(\lambda/\Lambda))r(F^{-1}(\lambda/\Lambda))} \quad \hat{\varepsilon}(\lambda) = \frac{\lambda\hat{\kappa}'(\lambda)}{\hat{\kappa}(\lambda)} = \frac{1}{\hat{\mu}(G^{-1}(1-\lambda/\Lambda))\hat{h}(G^{-1}(1-\lambda/\Lambda))},$$

where $r(t) = f(t)/F(t)$ is the reversed hazard rate of $F(t)$ and $\hat{h}(t) = g(t)/\bar{G}(t)$ is the failure rate of $G(t)$.

Proposition 2. *Let (t_L, t_H) $[(\hat{t}_L, \hat{t}_H)]$ represent a nondegenerate HiLo priority scheme in the fixed [increasing] values model.*

1. *If $\mu(t)$ is increasing, $\varepsilon(\lambda) > 1$ for all λ, and $\phi_{HL}(t_L, t_H) < \phi_{FF}(t_H)$.*

[2] The terminology for $\mu(t)$ is not standardized in the literature. Bagnoli and Bergstrom (2005) state (footnote 4) "We find it a bit surprising that our invidious civilization has not created a common English word for this idea, but we haven't been able to find such a word." before labeling it the "mean-advantage-over-inferiors." Other have termed it the "mean inactivity time" (Chandra and Roy, 2001).

2. *If* $\hat{\mu}(t)$ *is strictly decreasing,* $\hat{\varepsilon}(\lambda) > 1$ *for all* λ*, and* $\hat{\phi}_{HL}(t_L, t_H) < \hat{\phi}_{FF}(t_L)$.

3. *If* $\hat{\mu}(t)$ *is weakly increasing,* $\hat{\varepsilon}(\lambda) \leq 1$ *for all* λ*, and* $\hat{\phi}_{HL}(t_L, t_H) \geq \hat{\phi}_{FF}(t_L)$. *Equality holds only for the exponential distribution.*

Do priorities hurt customers? These results say, often times, yes. For either model of consumer utility, there exists a nontrivial class of distributions under which a move from FIFO to HiLo leaves customers worse off. A sufficient condition for a decreasing RMRL is that $F(t)$ has a decreasing reversed hazard rate, a commonly imposed – and commonly satisfied – assumption (Bagnoli and Bergstrom, 2005). Indeed, an increasing RMRL is the most sensible class of distribution to consider in our setting since a nonnegative random variable cannot have a RMRL that is either everywhere decreasing or constant (Chandra and Roy, 2001).

For the increasing values model, the story is slightly more nuanced since now customers can either be harmed or helped by a move to priorities. The MRL of a nonnegative random variable can be everywhere increasing or decreasing. In particular, a sufficient condition for a decreasing [increasing] MRL is an increasing [decreasing] failure rate (Lai and Xie, 2006). Thus there are many distributions for which either condition can hold. (Note that our results for the increasing values case are in some ways generalizations of those in Gurvich et al., 2018.)

What the two utility models have in common is the elasticity of the delay weightings, $\varepsilon(\lambda)$ and $\hat{\varepsilon}(\lambda)$. The delay weightings $\kappa(\lambda)$ and $\hat{\kappa}(\lambda)$ are both increasing in λ but the question is how quickly they increase. When the weighting are elastic (i.e., greater than one), surplus falls when moving from FIFO to HiLo. There is some clear intuition for the fixed values case. Here, surplus is increasing in expected delays. Further, $\kappa(\lambda)$ increases so quickly that the consumers are better off when the maximum possible weighting (i.e., $\kappa(\lambda_F)$) is applied to the average delay (W_F). In the increasing values case, a rapidly increasing $\hat{\kappa}(\lambda)$ implies that the net weighting on the low priority expected delay (i.e., $\hat{\kappa}(\lambda_F) - \hat{\kappa}(\lambda_H)$) is relatively large, which makes the move to HiLo unattractive.

Recall that a move FIFO to HiLo *always* increases social welfare. This implies that revenue gains swamp any decrease in consumer surplus. This does not guarantee that the drop in surplus is small. Figure 1 shows the percentage drop in surplus can be significant. The left-hand panel covers the fixed values case while the right-hand panel covers the increasing values case. Both panels show the ratio of surplus under HiLo to surplus on FIFO (i.e., $\phi_{HL}(t_L, t_H)/\phi_{FF}(t_H)$ or $\hat{\phi}_{HL}(\hat{t}_L, \hat{t}_H)/\hat{\phi}_{FF}(\hat{t}_L)$) for a set of Weibull distributions that differ in their shape parameter k. The fraction of customers opting for low-class service varies from 0

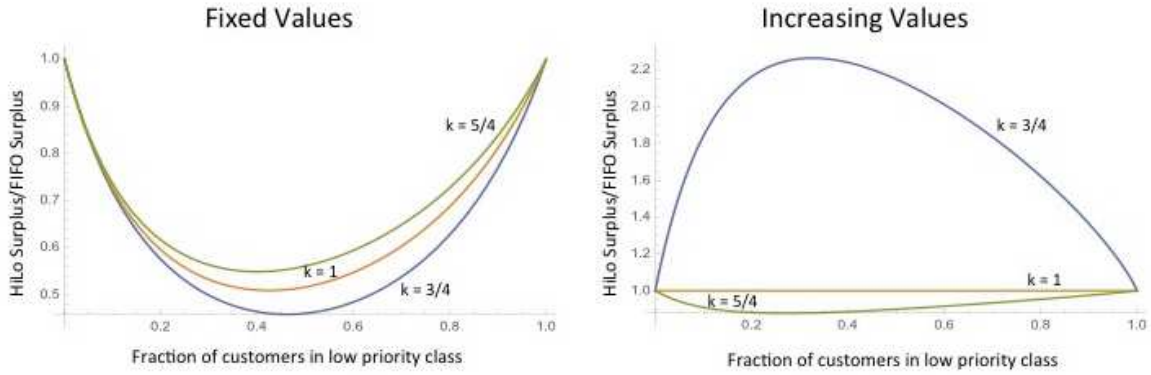

Figure 1: Consumer surplus under HiLo relative to consumer surplus under FIFO when types follow a Weibull distribution with scale parameter one and shape parameter k. All examples have a single server with a capacity of one customer per unit time, $\Lambda = 1$, and a utilization of 0.8. For the increasing values model, $V\left(\hat{t}\right) = \theta\left(k\right)\hat{t}^2$ where $\theta\left(k\right)$ is chosen such that $\bar{V}\left(\hat{t}_L\right) = 6$.

to 100%. The Weibull always has a decreasing RMRL; thus surplus always falls with a switch to priority service in the fixed values case. The Weibull has increasing MRL for shape parameters less than one but a decreasing MRL for shape parameters greater than one. For the special case of the exponential distribution (i.e., $k = 1$), the MRL is constant. The right-hand panel thus illustrates all three possible outcomes for the increasing values case.

For either utility model, the scale of the change in surplus depends on both how customers are split between the classes of service as well as the distribution. Obviously, if nearly all customers end up in one class, the system offers the majority of customers something very close to FIFO service and the gain or loss from the change in service discipline is small. When both classes see a substantial flow of customers, however, the impact can be large. All of the examples here have changes of at least 10% and some have changes in excess of 40%.

5.1. The role of variability

Characteristics of the type distribution also determine how a change in the service distribution affects customers. In the fixed values case (left-hand panel of Figure 1), a lower shape parameter results in a more significant decrease all else being equal. This is reversed in the increasing values case. A low shape parameter results in customers benefiting from a move to HiLo while a large shape parameter corresponds to a drop in surplus. For the Weibull distribution, an increasing shape parameter implies a lower coefficient of variation. These examples thus suggests that customers benefit from high (relative) variability in the increasing values case but are better off with lower variability in the fixed values case.

We now formalize this observation. Consider to random variables X_1 and X_2 with respective distributions $\Psi_1(x)$ and $\Psi_2(x)$. We say X_1 is smaller than X_2 in the convex transform order (and write $X_1 \preceq_c X_2$) if $\Psi_2^{-1}(\Psi_1(x))$ is convex for all x in the support of $\Psi_1(x)$ (Shaked and Shanthikumar, 2007). For $1 > \alpha > \beta > 0$, let $\phi_{HL}^i(\beta,\alpha) = \phi_{HL}\left(\Psi_i^{-1}(\beta), \Psi_i^{-1}(\alpha)\right)$ and $\hat{\phi}_{HL}^i\left(\Psi_i^{-1}(\beta), \Psi_i^{-1}(\alpha)\right)$ for $i = 1, 2$. Define $\phi_{FF}^i(\alpha)$ and $\hat{\phi}_{FF}^i(\beta)$ analogously.

Proposition 3. *Suppose $X_1 \preceq_c X_2$, then ...*

1. $\phi_{HL}^1(\beta,\alpha)/\phi_{FF}^1(\alpha) \geq \phi_{HL}^2(\beta,\alpha)/\phi_{FF}^2(\alpha)$.
2. *If $\hat{V}\left(\hat{t}\right) = \theta\hat{t}$, $\hat{\phi}_{HL}^2(\beta,\alpha)/\hat{\phi}_{FF}^2(\beta) \geq \hat{\phi}_{HL}^1(\beta,\alpha)/\hat{\phi}_{FF}^1(\beta)$.*

The convex transform order is a variability order and, in particular, implies a ranking of coefficients of variation (i.e., if $X_1 \preceq_c X_2$, X_1 has a lower coefficient of variation). The comparisons here are set up such that the arrival rates and thus expected waits are the same. Normalizing surplus under HiLo by the surplus under FIFO as well as assuming a linear value function, puts results under the two distributions on the same scale. In particular, the mean of the type distribution does not matter and the role of variability is highlighted.

5.2. The impact on individual customers

Focusing on consumer surplus demonstrates the impact of implementing priorities on the average customer. How a change in service discipline affects a particular, individual customer may differ. Consider the increasing values case. A customer with realized type $\hat{t}_L$ is indifferent to seeking service under either discipline by

construction. A customer with a slightly higher type gets a positive surplus under both regimes but prefers FIFO. Under HiLo, she has a longer wait and pays a lower price but that lower price is set to make a more patient customer (i.e., the one at $\hat{t}_L$) whole. This argument extends to show that any customer between $\hat{t}_L$ and $\hat{t}_H$ would be better off under FIFO.

At the other extreme, customers with extremely high waiting costs may be better off under HiLo. Consider a customer at $\hat{t}' >> \hat{t}_H$. This customer enjoys a shorter wait under HiLo and pays a higher price but that price is set to leave a customer at $\hat{t}_H$ indifferent between the two classes of service. The out of pocket price thus leaves the customer at $\hat{t}'$ with additional rents. If the support of the type distribution is unbounded, there must exist a $\hat{\tau} > \hat{t}_H$ such that all customers with realized types greater than $\hat{\tau}$ strictly prefer HiLo to FIFO. Moving from FIFO to HiLo would then create winners and losers; it cannot be the case that all customers are made better off or that all customers are worse off by the change in discipline. The fact that consumer surplus can increase or decrease with a move from FIFO to HiLo can then be seen as a reflection of the fact that different distributions place different waits on extreme ends of the range of possible waiting costs.

With fixed values, customers in the high priority class are certain to be hurt by a move from FIFO to HiLo. These customers enjoy shorter waits but must pay a price that is set to leave a more delay-sensitive customer (i.e., one at t_H) indifferent. The cost increase overvalues the waiting cost reduction for customers strictly between t_H and t_L. Low-priority customers could potentially benefit from the move to HiLo from FIFO. Some have waiting costs near zero and so are virtually indifferent to an increased delay as long as the price of the service falls relative to FIFO. Somewhat remarkably, the price does not necessarily fall.

Proposition 4. *In the fixed values setting, $p_L \geq p_F$ if $F(t)/t$ is decreasing.*

Under fixed-values, it may be that *all* customers are worse off under priorities. High priority customers receive better service but are overpaying for it. Still, they may be better off than low-priority customers who potentially endure longer waits while paying more than they would under FIFO. $F(t)/t$ will be decreasing if $f(t)$ is decreasing. Since all decreasing failure rate distributions have a decreasing density, there is a nontrivial set of distributions which satisfies the condition of the proposition. Finally, if types follow a uniform distribution, we must have that $p_L = p_F$.

5.3. Summary

Whether and to what extent the imposition of priorities lowers consumer surplus depends on the model of customer utility and on the distribution of customer types. Under fixed priorities, surplus falls when moving from FIFO to HiLo if the distribution has a decreasing reverse mean residual life (arguably, the most relevant case to consider). In this setting, variability hurts customers as the proportional loss from moving to HiLo is higher under a more variable distribution (as measured by the convex transform order. Finally, it is possible that all customers are worse off when priorities are implemented as the price for low priority service might be higher than the price under FIFO.

This last issue never arises under the increasing values model; the price of low priority service is always less than the FIFO price. Further, those with very high waiting costs (and hence very high values for the service) can be strictly better off under HiLo. The overall impact on customers consequently depends on the distribution of customer types. If the type distribution has an increasing MRL, consumer surplus is higher under HiLo than under FIFO. If, however, the type distribution has a decreasing MRL, surplus falls. Note that a sufficient condition for an increasing [decreasing] MRL is a decreasing [increasing] failure rate and that a decreasing [increasing] failure rate distribution always has a coefficient of variation greater [less] than one. Hence, it is not too surprising that under the increasing values model variability helps customers.

Finally, there is a link between the model settings in which surplus falls: the elasticity of the delay weightings. Under both models, the impact of the expected delay depends on the volume of customers served. When this impact is highly sensitive to volume (in the sense of having an elasticity greater than one), surplus falls when priorities are introduced.

6. Extensions

We now consider several extensions of our basic model. We first present an example in the increasing values setting with a type distribution that has a non-monotone MRL function. We next show how our results generalize to priority schemes with more than two levels of service. The results above assume that the total volume of customers served does not change as we move from FIFO to HiLo. We consequently also explore how things change when coverage is also in play. Finally, we present a generalization of the fixed values model in which customers

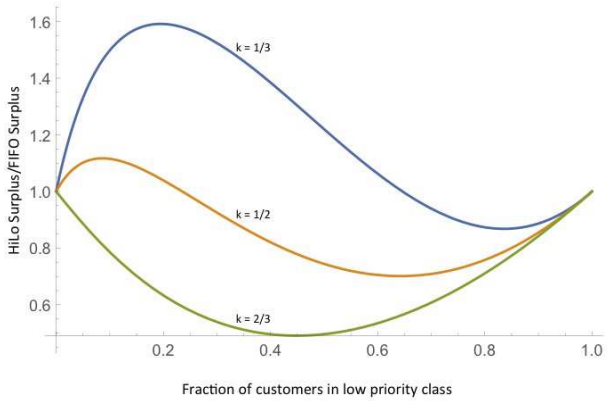

Figure 2: Increasing values with a non-monotone residual life function. Consumer surplus under HiLo relative to consumer surplus under FIFO when types follow a power function distribution with scale parameter nine and shape parameter k.

draw a valuation for the service that is independent of their waiting cost.

6.1. A non-monotone residual life function

Our results for the increasing values model have focused on monotone MRL functions. This allows for broad statements that hold for all possible HiLo schemes but ignores some common distributions that have non-monotone MRLs. The complication that follows from a non-monotone MRL is that we must consider just how customers are split between the high and low classes. From the proof of Proposition 2, surplus will be higher under FIFO if $\hat{\mu}\left(\hat{t}_L\right) > \hat{\mu}\left(\hat{t}_H\right)$ and higher under HiLo otherwise. Holding coverage – and hence $\hat{t}_L$ fixed – a non-monotone MRL will result in the value of $\hat{\mu}\left(\hat{t}_H\right)$ relative to $\hat{\mu}\left(\hat{t}_L\right)$ varying as the size of the high priority class changes.

Consider a power function distribution, i.e., $G(t) = (t/\Omega)^k$ for $0 \leq t \leq \Omega$ and $k > 0$. When the shape parameter k is less than one, the MRL is first increasing and then decreasing. Further the range over which the MRL is increasing is decreasing in k. If the MRL is increasing at $\hat{t}_L$, HiLo will increase consumer surplus for values of $\hat{t}_H$ close to $\hat{t}_L$ (i.e., for settings in which most customers are classified as high priority). However, as $\hat{t}_H$ moves further away from $\hat{t}_L$, $\hat{\mu}\left(\hat{t}_H\right)$ drops below

$\hat{\mu}\left(\hat{t}_L\right)$ and surplus is higher under FIFO. This is illustrated in Figure 2 which plots ratio of surplus under HiLo to surplus on FIFO (i.e., $\hat{\phi}_{HL}\left(\hat{t}_L, \hat{t}_H\right) / \hat{\phi}_{FF}\left(\hat{t}_L\right)$) as a function of what fraction of customers are in the low priority class for several power function distributions.

Before leaving this example, note that the coefficient of variation of the power function distribution is decreasing in k. Hence, we see that reduced variability in the increasing values case again hurts customers even when the MRL is not monotone.

6.2. More than two priority classes

To this point, we have only considered priority schemes with two levels of service. We now examine static priority schemes with $K > 2$ service classes. Assume that class $i+1$ has priority over class i. For the fixed values, let $T_K = (t_1, ..., t_K)$ be a vector of cutoff values corresponding to customers with types between t_i and t_{i-1} buying class i service where we follow the convention that $t_0 = 0$. Similarly define $\hat{T}_K = \left(\hat{t}_1, ..., \hat{t}_K\right)$ for the increasing values case. Here types between $\hat{t}_i$ and $\hat{t}_{i+1}$ purchase class i service where $t_{K+1} = \Omega$. We say that T_K or $\hat{T}_K$ represents a nondegenerate priority scheme if the arrival rate to each of the K classes is strictly positive.

With a slight abuse of notation, we can write consumer surplus for the fixed values case as

$$\phi_{HL}\left(T_K\right) = \Lambda \sum_{i=1}^{K} \left(\eta\left(t_i\right) - \eta\left(t_{i-1}\right)\right) W_i.$$

For the increasing values case we have

$$\hat{\phi}_{HL}\left(\hat{T}_K\right) = \Lambda \left[\bar{V}\left(\hat{t}_1\right) - \sum_{i=1}^{K} \left(\hat{\eta}\left(\hat{t}_i\right) - \hat{\eta}\left(\hat{t}_{i+1}\right)\right) W_i.\right].$$

Proposition 5. *Let T_K $\left[\hat{T}_K\right]$ be a nondegenerate priority scheme.*

1. *If the reversed hazard rate of $F(t), r(r)$, is decreasing, $\phi_{HL}\left(T_K\right) < \phi_{FF}\left(t_K\right)$.*
2. *If the failure rate of $G(t), \hat{h}(t)$, is strictly increasing, $\hat{\phi}_{HL}\left(\hat{T}_K\right) < \hat{\phi}_{FF}\left(\hat{t}_K\right)$.*
3. *If $\hat{h}(t)$ is decreasing, $\hat{\phi}_{HL}\left(\hat{T}_K\right) \geq \hat{\phi}_{FF}\left(\hat{t}_K\right)$. Equality holds for the exponential distribution.*

Our results for the HiLo case thus extend to less coarse priority schemes although we require slightly stricter assumptions. A decreasing reversed hazard rate implies an increasing RMRL while an increasing [decreasing] failure rate implies a decreasing [increasing] MRL. Consequently, we have that the loss (or gain) consumers see from the imposition of priorities is not due to the basic HiLo structure; it is simply a consequence of priorities.

The first two parts of the proposition do *not* imply that customers would always prefer a queue with J classes of service to one with $K > J$ classes. Such results cannot hold. If one sends all but a sliver of customer to, say, class 1 and split the rest equally among the other $K-1$ classes, surplus will be vary close to FIFO (recall Figure 1). Consumers would be better off under that scheme than under a HiLo scheme that equally splits customers between the two classes. A similar argument would show that the third part does not say that customers would always prefer a scheme with more classes if $G(t)$ has a decreasing failure rate.

6.3. Expanding coverage

Our comparisons between FIFO and HiLo have so far been predicated on coverage being fixed. That is, the total volume of customers being served is constant between the two system. A move from FIFO to HiLo, however, may also involve an increase in number of customers served (Afèche and Mendelson, 2004). Those additional customers may boost consumer surplus sufficiently to offset any loss in moving from FIFO from HiLo. We explore this point by considering how a social planner would manage the system. Specifically, we determine how many customers the planner would admit under FIFO and compare consumer surplus under this policy to the surplus achieved under the social welfare maximizing HiLo policy.

We begin with the increasing values model and focus on distributions with decreasing MRLs. Increasing throughput in this setting imposes a tradeoff even under FIFO. The expected value created gross of delay costs $\bar{V}(t)$ increases as t is reduced and the volume of customers goes up but delay costs also increase. Hence, there is a cutoff that would maximize consumer surplus. Since the social planner generally admits fewer customers than would maximize consumer surplus, there is room for the planner to increase surplus through increased volume.[3]

[3]Maximizing social welfare under FIFO would require $\hat{\phi}'_{FF}(t) = -\hat{\rho}'_{FF}(t)$. Since revenue maximization typically results in serving fewer customer than is socially optimal, we have that

k	$\left(\hat{\lambda}_L^* + \hat{\lambda}_H^*\right)/\hat{\lambda}_F^*$	$\hat{\phi}_{HL}/\hat{\phi}_{FF}$
1.0	116.3%	105.5%
1.5	110.4%	100.8%
2.0	107.7%	98.0%
2.5	106.0%	96.1%

k	$(\lambda_L^* + \lambda_H^*)/\lambda_F^*$	ϕ_{HL}/ϕ_{FF}
0.25	111.8%	134.7%
0.50	112.0%	122.1%
1.00	108.4%	112.1%
2.00	104.6%	105.6%

Table 1: Expanding coverage to maximize social welfare. Left-hand panel cover the increasing values case with $\hat{V}(t) = 3t$, Right-hand panel covers the fixed values case with $V = 5$. Throughout the mean is 10 and $\Lambda = 1$.

The left-hand panel of Table 1 presents examples in which types follow Weibull distributions with shape parameter k. We use stars to denote optimal values and compare the total throughput under HiLo ($\hat{\lambda}_L^* + \hat{\lambda}_H^*$) with the throughput under FIFO ($\hat{\lambda}_F^*$) as well as consumer surplus under the two schemes. A shape parameter of one corresponds to an exponential distribution which has a constant MRL. Consumers are then indifferent between FIFO and HiLo for a fixed level of coverage. Here, the move to HiLo increases throughput so consumer surplus increases. For larger values of k, surplus falls when priorities are implemented under fixed coverage. If the increase in throughput is sufficiently large (as it is for $k = 1.5$), consumers can still come out ahead when the social planner implements priorities. For higher values of the shape parameter, things are not as sanguine as consumer surplus falls as social planner implements priorities. There are two factors in play. First, the increase in volume is smaller and its benefit is swamped by the loss from moving to priorities. Second, from Proposition 3**,** we have that reduced variability results in a more significant loss in moving from FIFO to HiLo.[4]

The right-hand panel of Table 1 covers a similar example for the fixed values model. Here over a range of shape parameters (and thus a range of coefficients of variation), surplus increases with a move from FIFO to HiLo. Customer volume plays a different role under the fixed values model than it does under increasing values. Consider what happens under FIFO. In this setting, there is no tradeoff; surplus always increases with higher volume. Under HiLo, an increase in volume is also certain to increase consumer surplus if all of the additional customers go into the higher priority class. This does not guarantee, however, that consumer surplus

$\hat{\phi}'_{FF}(t) > 0$ at the socially optimal quantity.

[4]Compare these results to Gurvich et al. (2018) which examines how revenue maximization affects surplus.

Λ	$\frac{\lambda_L^*+\lambda_H^*}{\lambda_F^*}$	$\frac{\phi_{HL}}{\phi_{FF}}$	$\frac{p_H}{p_F}$	$\frac{p_L}{p_F}$
1.00	104.6%	105.6%	71.3%	53.5%
1.25	115.0%	130.8%	100.3%	81.2%
1.35	114.0%	109.6%	120.6%	102.9%
1.50	111.3%	77.3%	135.6%	122.3%
1.75	107.6%	47.1%	133.7%	126.0%

Table 2: Expanding coverage under increasing market size

will always go up as the social planner moves from FIFO to HiLo. Suppose that the overall market is large (i.e., Λ is significantly higher than available capacity). Then even under FIFO the social planner can operate at a very high utilization since there will be a large number of customers with low waiting costs. Because delay costs are low, the price can approach V even when waits are long. As seen from Table 2 in which the shape parameter is fixed at 1/4 and Λ is varied, it may be the case that prices for both classes of service are higher under HiLo than they were under FIFO.

In summary, a move from FIFO to HiLO results in the social planner serving additional customers, and this expansion of coverage may offset the reduction in consumer surplus that follows a move to HiLo under fixed coverage. Whether this happens under increasing values depends in part on the type distribution. When variability is relatively low, the increase in coverage is small and the loss from implementing priorities is high. Hence, surplus still falls. Under fixed values, the load on the system plays an important role. When the size of the market is large relative to capacity, there are many customers with very low waiting costs. HiLo allows the planner push the price of both classes of service close to V and surplus falls despite the increase in customers served.

6.4. Random valuations

As discussed above, the fixed valuations model is a simplification of a world in which customers are defined by a two dimensional type, a random per-unit-time waiting cost as well as a random and independently drawn valuation for the service. We now consider such a setting. Waiting costs are still drawn from a continuous distribution $F(t)$ but that valuations are either V^+ with probability δ or V^- with probability $1-\delta$ for $V^+ > V^- > 0$ for δ fixed and independent of t.

Our goal is to compare consumer surplus under FIFO and HiLo assuming the

same total volume of customers is served under the two service disciplines. In the fixed values case, this is straightforward since the same set of customers is served as we move from FIFO to HiLo. The complication that arises in the random values setting is that setting prices and wait to yield the same volume of customer under both FIFO and HiLo results in a different set of customers being served.

In what follows, we focus on a setting in which the overall arrival rate is $\alpha\Lambda$ for some α between zero and one. Under HiLo, we assume that the arrival rate for the low class is $\beta\Lambda$ for some β between zero and α. Let $t_\alpha = F^{-1}(\alpha)$ and $t_\beta = F^{-1}(\beta)$. With fixed values, all customer with realized types less than t_α would patronize the system. Under HiLo, those between zero and t_β would join the low priority class while those between t_β and t_α would join the high priority class. This simple structure does not hold under random values. In the FIFO case, if a customer at t_α with valuation V^+ is indifferent to buying given price p_F and expected wait W_F, a customer with the same waiting cost but valuation V^- strictly prefers not buying. The FIFO regime consequently requires two cutoff values, t_F^+ and t_F^-. The former is the waiting costs at which a customer with valuation V^+ is indifferent to joining the system while t_F^- is the corresponding value for those with valuation V^-. The pair t_F^+ and t_F^- is unique and $t_F^- < t_\alpha < t_F^+$.

The HiLo case similarly requires an extra cutoff. Let t_H^+ be the type indifferent to purchasing given a valuation of V^+ and t_L^+ be the type who is indifferent between the two classes of service given a valuation of V^+. The third cutoff t_{HL}^- determines the customer with valuation V^- who is indifferent to buying. The raises the question of which class of service a customer at t_{HL}^- would buy. There are two cases. In Case 1, $F\left(\frac{V^+-V^-}{W_H} + t_\beta\right) \leq \frac{\alpha-\beta}{\delta} + \beta$; a customer at t_{HL}^- buys high priority service and $t_L^+ = t_\beta$. In Case 2, $F\left(\frac{V^+-V^-}{W_H} + t_\beta\right) > \frac{\alpha-\beta}{\delta} + \beta$ and $t_{HL}^- < t_\beta$. Customers with valuation V^- only buy low priority service, and we must have $t_L^+ > t_\beta$.

In both cases, $t_H^+ > t_\alpha > t_{HL}^-$. Additionally, one can show that $t_H^+ - t_{HL}^- > t_F^+ - t_F^-$, which leads to a shift in the mix of the customer served. HiLo induces more high waiting cost customers to buy while dissuading some lower cost customers from purchasing. The overall impact is ambiguous. Reasoning from the intuition developed in the fixed values model, consumer surplus increases as the gap between the marginal customer and the average customer grows. Here HiLo pushes the marginal customer out relative FIFO but also leaves out some customers who are relatively far from the marginal customer.

Turning to how priorities affect consumer surplus, Table 3 presents examples

k	p_H	p_L	p_F	$\frac{\phi_{HL}}{\phi_{FF}}$
0.50	48.60	37.35	14.93	56.0%
0.75	50.62	28.57	18.86	62.1%
1.00	52.50	22.50	22.50	67.4%
1.25	53.99	18.36	25.34	71.6%
1.50	55.16	15.48	27.55	74.9%

k	p_H	p_L	p_F	$\frac{\phi_{HL}}{\phi_{FF}}$
0.50	73.91	62.11	49.47	80.1%
0.75	81.91	56.84	53.62	79.4%
1.00	86.88	52.50	57.50	80.0%
1.25	90.11	49.45	60.61	80.6%
1.50	92.37	47.37	63.09	80.9%

Table 3: Comparing surplus under random values. The left-hand panel represents Case 1 with $\mathbf{V}^+ - \mathbf{V}^- = \mathbf{10}$. The right-hand panel represents Case 2 with $\mathbf{V}^+ - \mathbf{V}^- = \mathbf{50}$. Additionally, $\Lambda = 1$,$\alpha = 0.75$, $\beta = 1/3$.

for both Case 1 (left-hand panel) and Case 2 (right-hand panel). In all examples, types follow a power function distribution with a mean of ten. That is, $F(t) = \left(\frac{t}{\Omega(k)}\right)^k$ for $t \in (0, \Omega(k))$ where $\Omega(k) = 10 \times (1 + 1/k)$. The lower value V^- is fixed in all examples at 75 and δ is always 0.75. However, V^+ varies between the two cases. For the first case, $V^+ = 85$ but in the second $V^+ = 125$. Several insights are apparent. First, priorities continue to hurt customers. Second, we may have that all customers pay more out of pocket under priorities than under FIFO. $F(t)/t$ is decreasing for $k < 1$ but is constant for $k = 1$. For Case 1, the results here are consistent with the fixed values case with $p_L > p_F$ for $k < 1$ but equal for $k = 1$. (Indeed, one can show for uniformly distributed types that $p_L = p_F$ in Case 1.) In Case 2, however, we have $p_L > p_F$ even for $k = 1$. Finally, note that a lower k corresponds a higher coefficient of variation. Case 1 is then consistent with the fixed values case and customers incur a higher penalty from the switch with greater variability. Case 2 is different and the hit customers take is essentially independent of the variability in the types.

7. Conclusion

We have examined a queue serving delay-sensitive customers who are privately informed of their waiting costs. Managing the queue thus requires posting expected waits and out-of-pocket prices that induce the appropriate set of customers to buy a particular class of service. In this setting, we have presented conditions such that a move from FIFO service to priority service systematically benefits or harms consumers. The outcome depends on customer utility model and the distribution of customer types. If all customers have the same value for the service, then

priorities reduce consumer surplus for essentially all relevant type distributions. Additionally, the loss customers experience increases as the type distribution becomes more variable. If customers with higher waiting costs have a higher value for the service, consumer surplus falls with a move to priorities if the type distribution has a decreasing MRL. If the MRL is increasing, surplus is higher under priorities. With increasing values, customers are better off with a more variable type distribution.

Our goal here is not to argue that queues should be managed to maximize consumer surplus. Rather, we have shown that priorities can impact consumers in systematic – often negative – ways. At face value, priorities would always seem to be good. Consumers have more choices and social welfare increases relative to serving customers in the order of their arrival. That welfare increase may mask that consumers are actually hurt when priorities are used instead of FIFO. Indeed, we show it is possible that everyone is wore off under priorities as even those given low priorities pay higher prices than they would under FIFO. Implementing priorities may expand the number of customers served but that does not necessarily guarantee that consumer surplus will increase.

As noted, our fixed values model is a simplified version of a setting in which service values are drawn independently of waiting costs. Our increasing values model is a simplification of a world in which waiting costs and customer values are positively correlated. We have presented an example showing that the intuition from our fixed values model is relevant when valuations are random but independent of waiting costs. An extension of this work would be consider having valuations drawn randomly but in a way correlated with waiting costs. While there is only one way to have independent valuations, there are many ways to model positive correlation. For at least some of these, the results of the *fixed* values model may in fact be relevant. For example, suppose that the support of values is fixed but the distribution of values is stochastically increasing in the customer's type. That gives positive correlation between waiting costs and valuations. However, for any given value realization there will be some maximum waiting cost that leads to purchase. That is, which customers buy will be determined by an upper bound on the waiting cost as in our fixed values model. Thus there is reason to believe that our broad results from the fixed values model could still apply if types and values are correlated.

References

Afèche P (2013) Incentive-compatible revenue management in queueing systems: Optimal strategic delay. Manufacturing Service Oper. Management 15(3):423–443.

Afèche P, Mendelson, H (2004) Pricing and priority auctions in queueing systems with a generalized delay cost structure. Management Sci. 50(7):869–882.

Arriaza A, Sordo M. A., Súarez-Llorens A, (2017) Comparing residual lives and inactivity times by transform stochastic orders. IEEE Transactions on Reliability. 66(2): 366–372.

Bagnoli M, Bergstrom T (2005) Log-concave probability and its applications. Econom. Theory 26(2):445–469.

Barnes, B (2013) At Theme Parks, a V.I.P. Ticket to Ride, New York Times (June 9) https://www.nytimes.com/2013/06/10/business/at-universal-park-a-vip-pass-to-help-lift-revenue.html.

Chandra N, Roy D (2001) Some Results on Reversed Hazard Rate. Probability in the Engineering and Informational Sciences, 15(1), 95-102.

Chen N, Guillermo G (2018) Welfare Analysis of Dynamic Pricing. Management Science, forthcoming.

Chen Y, Schwartz M. (2015). Differential pricing when costs differ: A welfare analysis. The RAND Journal of Economics, 46(2), 442-460.

Cowan S (2012) Third-degree price discrimination and consumer surplus. J. Indust. Econom. 60(2):333–345.

Cox D, Smith W (1961) Queues (Methuen & Co., London)

Doroudi S, Akan M, Harchol-Balter M, Karp J, Borgs C, Chayes JT (2013) Priority pricing in queues with a continuous distribution of customer valuations. Technical Report CMU-CS-13-109, Carnegie Mellon University, Pittsburgh.

Dupuit, J (1844). De la Measure de l'Utilité des Travaux Publiques. In Annales des Ponts et Chaussées (Vol. 8) (Translated and reprinted in: K. Arrow & T. Scitovski (Eds.) (1969). AEA Readings in Welfare Economics, AEA, pp. 255–283.)

Flandersnews.be (2013) Walibi to introduce queue jumpers pass. (October 6), http://deredactie.be/cm/vrtnieuws.english/Life/1.1651691.

Gavirneni N, Kulkarni V (2016) Self-selecting priority queues with Burr distributed waiting costs. Production Oper. Management 25(6):979–992.

Ghanem SB (1975) Computing center optimization by a pricing priority policy. IBM Systems J. 14(3):272–291.

Gilland WG, Warsing DP (2009) The impact of revenue-maximizing priority pricing on customer delay costs. Decision Sci. 40(1): 89–120.
Gurvich I, Lariviere, MA, Ozkan C (2018) Coverage, Coarseness, and Classification: Determinants of Social Efficiency in Priority Queues. Management Science, forthcoming.
Johnston, J (2104) Customers can pay 50p to avoid having to queue to speak to an operator, The Independent (August 15), https://www.independent.co.uk/lifestyle/gadgets-and-tech/news/ee-criticised-for-introducing-two-tier-call-centre-system-9670429.html.
Hassin R, Haviv M (2003) Queue or Not to Queue: Equilibrium Behavior in Queueing Systems (Kluwer Academic Publishers, Boston).
Larson, R (2012) Are priority queues 'un-American'? BBC News (October 10), http://www.bbc.com/news/magazine-19712847.Lai C-D, Xie M (2006) Stochastic Ageing and Dependence for Reliability (Springer Science and Business Media, New York).
Naor P (1969) The regulation of queue size by levying tolls. Econometrica 37(1):15–24
Puzzanghera, J (2017) Net neutrality's repeal means fast lanes could be coming to the internet. Is that a good thing?, Los Angeles Times (December 13), http://www.latimes.com/business/la-fi-net-neutrality-fast-lanes-20171213-story.html.
Schatz, A (2010) Patent Fast Track Proposed, Wall Street Journal (June 3), https://www.wsj.com/articles/SB10001424052748704515704575282951991894276
Shaked M, Shanthikumar JG (2007) Stochastic Orders (Springer, New York).
Stamatopoulos I, Chehrazi N, Bassamboo A (2017) Welfare Implications of Inventory-Driven Dynamic Pricing working paper, University of Texas at Austin.

Appendix

Proof of Proposition 1: For fixed values, we can rewrite our revenue functions as follows:

$$\begin{aligned}\rho_{HL}(t_L,t_H) &= \Lambda\left[F(t_H)V - F(t_H)t_H\left(W_H + \frac{F(t_L)t_L}{F(t_H)t_H}(W_L - W_H)\right)\right]\\ \rho_{FF}(t_H) &= \Lambda\left[F(t_H)V - F(t_H)t_H\left(W_H + \frac{F(t_L)}{F(t_H)}(W_L - W_H)\right)\right],\end{aligned}$$

where the expression for $\rho_{FF}(t_H)$ follows from (1). The result then follows from $t_L < t_H$. The increasing values case is similar but rewrites (1) as $W_F = W_L - \frac{\bar{G}(\hat{t}_H)}{\bar{G}(\hat{t}_L)}(W_L - W_H)$.

For social welfare, rewrite $\omega_{HL}(t_L,t_H)$ and $\omega_{FF}(t_H)$ as

$$\begin{aligned}\omega_{HL}(t_L,t_H) &= \Lambda\left[F(t_H)V - \int_0^{t_H} tf(t)\,dt\left(W_H + \frac{\int_0^{t_L} tf(t)\,dt}{\int_0^{t_H} tf(t)\,dt}(W_L - W_H)\right)\right]\\ \omega_{FF}(t_H) &= \Lambda\left[F(t_H)V - \int_0^{t_H} tf(t)\,dt\left(W_H + \frac{F(t_L)}{F(t_H)}(W_L - W_H)\right)\right].\end{aligned}$$

The result then follows because $\frac{\int_0^t tf(t)dt}{F(t)}$ is an increasing function. The increasing value case is similar and depends on $\frac{\int_t^\Omega xg(x)dx}{\bar{G}(\hat{t})}$ being an increasing function. □

Proof of Proposition 2: We write surplus under fixed values as:

$$\begin{aligned}\phi_{HL}(t_L,t_H) &= \Lambda\eta(t_H)\left[W_H + \frac{\eta(t_L)}{\eta(t_H)}(W_L - W_H)\right]\\ \phi_{FF}(t_H) &= \Lambda\eta(t_H)\left(W_H + \frac{F(t_L)}{F(t_H)}(W_L - W_H)\right).\end{aligned}$$

Surplus is then higher under FIFO if $\frac{F(t_L)}{F(t_H)} > \frac{\eta(t_L)}{\eta(t_H)}$, which holds if $\mu(t)$ is increasing. The elasticity result follows from the relationship $\mu(t)r(t) = 1 - \mu'(t)$. For increasing values, we have

$$\begin{aligned}\hat{\phi}_{HL}(\hat{t}_L,\hat{t}_H) &= \hat{\omega}_{HL}(\hat{t}_L,\hat{t}_H) - \hat{\rho}(\hat{t}_L,\hat{t}_H)\\ &= \Lambda\left[\bar{V}(t) - \hat{\eta}(\hat{t}_L)\left(W_L - \frac{\hat{\eta}(\hat{t}_H)}{\hat{\eta}(\hat{t}_L)}(W_L - W_H)\right)\right]\\ \hat{\phi}_{FF}(\hat{t}_L) &= \Lambda\left[\bar{V}(t) - \hat{\eta}(\hat{t}_L)\left(W_L - \frac{\bar{G}(\hat{t}_H)}{\bar{G}(\hat{t}_L)}(W_L - W_H)\right)\right].\end{aligned}$$

Surplus will be higher [lower] under FIFO if $\frac{\bar{G}\left(\hat{t}_H\right)}{\bar{G}\left(\hat{t}_L\right)} > [<] \frac{\hat{\eta}\left(\hat{t}_H\right)}{\hat{\eta}\left(\hat{t}_L\right)}$, which holds if $\hat{\mu}\left(\hat{t}\right)$ is decreasing [increasing]. The elasticity result follows from the relationship $\hat{\mu}\left(t\right)\hat{h}\left(t\right) = 1 + \mu'\left(t\right)$. □

Proof of Proposition 3: Define $\eta^i\left(t\right) = t\Psi_i\left(t\right) - \int_0^t x\psi_i\left(x\right)dx = \int_0^t \Psi_i\left(x\right)dx$, where the second relationship follows from an integrations by parts. Also note that in the fixed values case, the arrival rate to the low and the high class will be $\Lambda\beta$ and $\Lambda\left(\alpha - \beta\right)$, respectively. Arrival rate and expected waits are then the same under both type distributions. We next have

$$\begin{aligned}
\phi^i_{HL}\left(\beta,\alpha\right) &= \Lambda\eta^i\left(\Psi_i^{-1}\left(\alpha\right)\right)\left[W_H + \frac{\eta^i\left(\Psi_i^{-1}\left(\beta\right)\right)}{\eta^i\left(\Psi_i^{-1}\left(\alpha\right)\right)}\left(W_L - W_H\right)\right] \\
\phi^i_{FF}\left(\alpha\right) &= \Lambda\eta^i\left(\Psi_i^{-1}\left(\alpha\right)\right)W_F.
\end{aligned}$$

$\phi^1_{HL}\left(\beta,\alpha\right)/\phi^1_{FF}\left(\alpha\right)$ is thus greater than $\phi^2_{HL}\left(\beta,\alpha\right)/\phi^2_{FF}\left(\alpha\right)$ if $\frac{\eta^1\left(\Psi_1^{-1}(\xi)\right)}{\eta^2\left(\Psi_2^{-1}(\xi)\right)}$ is decreasing in ξ for $1 > \xi > 0$. This follows because the convex transform order implies the quantile mean inactivity time order (Arriaza et al., 2017).

For the increasing values case, we have $\hat{\eta}^i\left(t\right) = \int_t^\Omega x\psi_i\left(x\right)dx - t\bar{\Psi}_i\left(t\right) = \int_t^\Omega \bar{\Psi}_i\left(x\right)dx$ and note that the arrival rate to the high class is now $\Lambda(1-\alpha)$ and the low arrival rate is $\Lambda\left(\alpha - \beta\right)$. Also, the linearity of $\hat{V}\left(t\right)$ implies $\bar{V}^i\left(t\right) = \theta\hat{\eta}^i\left(t\right)$.

$$\begin{aligned}
\hat{\phi}^i_{HL}\left(\beta,\alpha\right) &= \Lambda\hat{\eta}^i\left(\Psi_i^{-1}\left(\beta\right)\right)\left[\theta - W_L + \frac{\hat{\eta}^i\left(\Psi_i^{-1}\left(\alpha\right)\right)}{\hat{\eta}^i\left(\Psi_i^{-1}\left(\beta\right)\right)}\left(W_L - W_H\right)\right] \\
\hat{\phi}^i_{FF}\left(\beta\right) &= \Lambda\hat{\eta}^i\left(\Psi_i^{-1}\left(\beta\right)\right)\left(\theta - W_F\right).
\end{aligned}$$

$\hat{\phi}^1_{HL}\left(\beta,\alpha\right)/\hat{\phi}^1_{FF}\left(\beta\right)$ is thus less than $\hat{\phi}^2_{HL}\left(\beta,\alpha\right)/\hat{\phi}^2_{FF}\left(\beta\right)$ if $\frac{\hat{\eta}^2\left(\Psi_2^{-1}(\xi)\right)}{\hat{\eta}^1\left(\Psi_1^{-1}(\xi)\right)}$ is increasing in ξ for $1 > \xi > 0$. This follows because the convex transform order implies the DMRL order (Shaked and Shanthikumar, 2007). □

Proof of Proposition 4: We have $p_L = V - t_H\left(W_H + \frac{t_L}{t_H}\left(W_L - W_H\right)\right)$ and $p_F = V - t_H\left(W_H + \frac{F(t_L)}{F(t_H)}\left(W_L - W_H\right)\right)$. p_L is then higher if $\frac{F(t_L)}{F(t_H)} > \frac{t_L}{t_H}$, which holds if $\frac{F(t)}{t}$ is increasing. □

Proof of Proposition 5: We begin with the following lemma.

Lemma 6. *If $r\left(t\right)$ is decreasing, $\kappa\left(\lambda\right)$ is convex. If $\hat{h}\left(t\right)$ is increasing [decreasing], $\hat{\kappa}\left(\lambda\right)$ is concave [convex].*

Proof: Differentiating, we have

$$\begin{aligned}\kappa'(\lambda) &= \frac{\lambda/\Lambda}{\Lambda f\left(F^{-1}(\lambda/\Lambda)\right)} = \frac{1}{\Lambda r\left(F^{-1}(\lambda/\Lambda)\right)} \\ \hat{\kappa}'(\lambda) &= \frac{\lambda/\Lambda}{\Lambda g\left(G^{-1}(1-\lambda/\Lambda)\right)} = \frac{1}{\Lambda \hat{h}\left(G^{-1}(1-\lambda/\Lambda)\right)}.\end{aligned}$$

If $r(t)$ is decreasing, $\kappa'(\lambda)$ is increasing and $\kappa(\lambda)$ is convex. The argument for $\hat{\kappa}(\lambda)$ is similar. □

The increasing values case is similarly proven in Lemma 3 of Gurvich et al. (2018). Note that the lemma implies that $\Delta(t) = \frac{\eta(t)-\eta(t')}{F(t)-F(t')}$ is increasing in t for $t > t'$. The proof for the fixed values case then proceeds by induction. Suppose the result holds for $K-1$ classes and consider a queue managed with K classes. For $i < K-1$, let W_i and W_{i+1} be the expected waits for classes i and $i+1$, respectively. Their contribution to consumer surplus is:

$$\begin{aligned}&\Lambda\left[W_i\left(\eta(t_i) - \eta(t_{i-1})\right) + W_{i+1}\eta(t_{i+1}) - \eta(t_i)\right] \\ =\ &\Lambda\left(\eta(t_{i+1}) - \eta(t_{i-1})\right)\left[W_{i+1} + \frac{\eta(t_i) - \eta(t_{i-1})}{\eta(t_{i+1}) - \eta(t_{i-1})}\left(W_i - W_{i+1}\right)\right].\end{aligned}$$

Suppose we collapse classes into one class. Note that this will have no impact on any customer outside of these classes. The expected wait for customers with values between t_{i-1} and t_{i+1} is now

$$\bar{W} = \frac{F(t_i) - F(t_{i-1})}{F(t_{i+1}) - F(t_{i-1})}W_i + \frac{F(t_{i+1}) - F(t_i)}{F(t_{i+1}) - F(t_{i-1})}W_{i+1}$$

and their contribution to consumer surplus is then

$$\Lambda\left(\eta(t_{i+1}) - \eta(t_{i-1})\right)\left(W_{i+1} + \frac{F(t_i) - F(t_{i-1})}{F(t_{i+1}) - F(t_{i-1})}\left(W_i - W_{i+1}\right)\right).$$

Customers are then better off if classes i and $i+1$ are collapsed since $\Delta(t)$ is increasing.

For the increasing values case, $\frac{\hat{\kappa}(\lambda')-\hat{\kappa}(\lambda)}{\lambda'-\lambda}$ is decreasing in λ for $\lambda < \lambda'$ if $\hat{h}(t)$ is increasing. That implies that $\hat{\Delta}\left(\hat{t}\right) = \frac{\hat{\eta}\left(\hat{t}'\right)-\hat{\eta}\left(\hat{t}\right)}{\bar{G}\left(\hat{t}'\right)-\bar{G}\left(\hat{t}\right)}$ for $\hat{t} > \hat{t}'$ is increasing in $\hat{t}$ since a higher cutoff corresponds to lower volume. The proof is then parallels the fixed values case. □